

THE GENE STRUCTURE AND THE POLYMORPHISM OF
THE HUMAN COMPLEMENT COMPONENT C4

Chack-yung Yu, B.Sc., M.Phil.

Trinity College, Oxford

A thesis submitted for the
Degree of Doctor of Philosophy

October 1987



MRC Immunochemistry Unit,
Department of Biochemistry,
University of Oxford.

Abstract

The Gene Structure and the Polymorphism of the Human Complement Component C4

Chack-yung Yu

Submitted for the degree of D. Phil.

Trinity College, Oxford

Trinity term, 1987

1. The DNA sequence of the human complement C4A gene from a cosmid clone Cos 3A3 was determined and the complete exon-intron structure elucidated. The 5' flanking region of the C4 gene contains three TATA sequences and a transcriptional enhancer core sequence, which are >200 nucleotides (nt) and 60-70 nt upstream from the CAP site, respectively. The gene consists of 42 exons coding for a precursor protein of 1745 residues. The first exon codes for a 51 nt 5' untranslated sequence, a leader peptide of 19 residues, and the N-terminus of the β chain. The β - α and the α - γ chain junctions are encoded by exons 17 and 34, respectively. The anaphylatoxin C4a and the thiolester site are encoded by phase 1-1 symmetrical exons. Most of the amino acids encoded at the splice junctions are polar or charged. Between exons 10 and 11 is a 6-7 kb intron that is flanked by direct long terminal repeats and may be absent in some C4 genes located at the second C4 locus. The last exon codes for the C-terminus of the γ chain and a 140 bp 3' untranslated sequence. The intergenic region between the C4 gene and its neighbouring 21-hydroxylase (21OHase) gene is ~3028 bp.
2. Eighteen polymorphic amino acids on C4 have been identified through genomic DNA, cDNA and protein sequencing. Fourteen of them are located on the α chain (C4a: 2 changes; C4d: 12 changes). The rest are scattered on the β and the γ chains. There are potential size variations by one residue on the β chain, and by a tripeptide that contains a sulphation site on the α chain.
3. Four common and rare C4 alleles have been cloned from individuals whose C4 proteins were chemically and serologically characterised. Analysis of the sequences at the C4d regions has allowed the identification of the C4A/C4B isotypic residues at positions 1101-6: C4A has the sequence PCPVLD, while C4B has the sequence LSPVIH. Presumably these isotypic residues are the cause of the class-specific, differential chemical reactivities. Moreover, the probable locations for the two Rodgers (Rg) and the six Chido (Ch) antigenic determinants were deduced. The C4B isotypic residues may be involved in the expression of the Ch2 and the Ch4 epitopes, while the C4A isotypic residues may not be related to either of the Rg determinants.
4. Definitive restriction fragment length polymorphisms (RFLPs) representing the exact locations responsible for the isotypicity between C4A and C4B, and for their generally associated Rg1 and Ch1 antigenic determinants, have been designed. In combination with the Taq I polymorphic patterns specific for the C4 and for the 21OHase gene loci, it has been shown that the null allele of the HLA haplotype B44 DR6 C4A 3 C4B Q0 is not a C4B allele, but probably encodes another C4A 3 allotype at the second C4 locus.

Acknowledgements

I would like to dedicate this thesis to the memory of my teacher, the late Professor Rodney Porter, CH, FRS. His constant encouragements, subtle criticisms, simple and clear direction have built up my confidence and enlightened my desire for scientific pursuit. His sudden tragic accident was the deepest sorrow to me. The benevolent helps to me by Prof.'s friends and colleagues after his death are heartily acknowledged.

I am much grateful to Dr. R. Duncan Campbell, who succeeded Prof. as my supervisor, for invaluable helps and advice.

I would like to express my sincere gratitude to Tertia Belt who instructed me the techniques in molecular biology; to Mike Carroll who supported and appreciated my works; and to the Unit director Dr. Ken Reid who gave me encouragements in many occasions.

The lively atmosphere in the Unit is most memorable. I appreciated David Bentley for his wit and critical discussion, and Dr. Duncan Campbell and Ms. Sandra Smith for maintaining a very convenient and well-ordered laboratory.

I am indebted to Edith Sim, Bob Sim and Alex Law for helpful discussions; to Alister Dodds and Carolyn Giles (Royal Postgraduate Medical School, London) for fruitful collaborations; and to Ms Carolyn Brooks and Mr Frank Caddick for expert secretarial assistance.

This thesis would not have been finished without the love, help and understanding of my dear wife, Lai-chu and my parents in Hong Kong.

I gratefully acknowledge the financial support from the Commonwealth Scholarship Commission. I would also like to thank the Unit, the Trinity College, the Graduate studies, and the Edward P. Abraham Trust, the for providing financial assistance to me for attending scientific meetings.

ABBREVIATIONS

Standard abbreviation and symbols recommended by the IUPAC-IUB Commission on Biochemical Nomenclature have generally been used. Activated components are indicated by a bar, e.g. C $\bar{I}$. Non-standard abbreviations are described below.

a.a.	amino acid
α_2 M	α_2 -macroglobulin
bp	base pairs
BSA	bovine serum albumin
C4bp	C4b binding protein
cDNA	complementary DNA
Ch	Chido
CR1	complement receptor 1
DAF	decay accelerating factor
ddNTP	dideoxynucleosides triphosphate
dNTP	deoxynucleoside triphosphate
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid (disodium salt)
EM	electron microscopy
EtOH	ethanol
HLA	human leukocyte antigen
IgG, etc.	Immunoglobulin G, etc.
kb	kilobases
LHRs	long homologous repeating units
MAC	membrane attack complex
MCP	membrane cofactor protein
MHC	major histocompatibility complex

Mol. wt.	molecular weight
mRNA	messenger RNA
n.c.	non-coding
nt	nucleotide
NTP	ribonucleoside triphosphate (N = A, G, C or T)
OD	optical density
PEG	polyethylene glycol
PIPES	piperazine-N-N'-bis (2-ethane-sulphonic acid)
rATP	adenosine triphosphate
RFLP	restriction fragment length polymorphism
Rg	Rodgers
RNA	ribonucleic acid
SCO1	complotype: factor B S, C2 C, C4A QO, C4B 1
SCR	short consensus repeating units
SDS	sodium dodecyl sulphate
s.j.	splice junction
SLE	systemic lupus erythematosus
ssDNA	sonicated salmon sperm DNA
TEMED	N,N,N',N'-tetramethylethylenediamine
Tris	tris (hydroxymethyl) aminoethane

TABLE OF CONTENTS

ABSTRACT

ACKNOWLEDGEMENTS

Chapter 1 Introduction

1.1	General introduction	1
1.2	The complement system	3
1.2.1	Introduction	3
1.2.2	The classical pathway of activation	6
1.2.2.1	The C1 complex	6
1.2.2.2	Assembly of the C3 and C5 convertases	9
1.2.2.3	The regulation of the classical pathway C3 convertase	14
1.2.3	The alternative pathway of activation	20
1.2.4	The membrane attack complex	24
1.3	Correlation of the exon structure and protein structure/function	29
1.4	The complement component C4	35
1.4.1	Biochemistry of human C4	35
1.4.2	Genetics of human C4	41

Chapter 2 Materials and methods

2.1	Materials	44
2.1.1	Enzymes	44
2.1.2	Chemical reagents	44
2.1.3	Bacterial strains, plasmids and phages	45
2.1.4	Radioactive molecules	46
2.2	Radioactive molecules	46
2.3	Preparation of DNA	46
2.3.1	Small scale isolation of plasmid DNA	46

2.3.2	Large scale isolation of plasmid DNA	47
2.3.3	Isolation of replicative form of bacteriophage M13 mp9	48
2.3.4	Small scale isolation of bacteriophage DNA	49
2.3.5	Large scale isolation of bacteriophage DNA	50
2.3.6	Isolation of human chromosomal DNA	50
2.4	Restriction digests of DNA	51
2.5	Radiolabelling of DNA	52
2.5.1	'Fill-in' reaction	52
2.5.2	Kinase labelling	52
2.5.3	Nick-translation	53
2.5.4	Multi-primed labelling	53
2.6	Gel eletrophoresis	53
2.6.1	Polyacrylamide gel eletrophoresis	53
2.6.2	Autoradiography	54
2.6.3	Agarose gel eletrophoresis	55
2.7	Recovery of DNA from gels	55
2.7.1	Elution of DNA from acrylamide gels	55
2.7.2	Elution of DNA from agarose gels	56
2.8	Subcloning of restricted DNA fragments into pAT 153 / Pvu II /8 vector	56
2.8.1	Preparation of <u>E. coli</u> MC1061 competent cells	56
2.8.2	Ligation and transformation	57
2.8.3	Screening of colonies for positives	57
2.9	Organic synthesis of oligonucleotides by the phosphotriester method	58
2.9.1	Support functionalisation and chain assembly	59

2.9.2	Deprotection and purification of oligonucleotides	60
2.10	Dideoxy DNA sequencing	61
2.10.1	Preparation of vector	61
2.10.2	Subcloning of DNA into M13 vectors	61
2.10.3	Preparation of phage DNA	63
2.10.4	Relative orientation of integrated fragments	64
2.10.5	Sequencing reaction	64
2.10.6	Electrophoresis of sequenced samples	65
2.10.7	Analysis of nucleotide sequence data	66
2.11	DNA transfer blot analysis	67
2.11.1	Analysis of DNA fragments between 0.5-15 kb in size	67
2.11.2	Analysis of DNA fragments between 0.1-1.5 kb in size	68
2.12	Construction of genomic libraries	69
2.12.1	Preparation of genomic DNA for cloning	69
2.12.2	Preparation of λ vector arms	70
2.12.3	Packaging extracts and packaging reactions	71
2.12.4	Ligation, Packaging, and assay of recombinants	71
2.12.5	Plating and screening of λ EMBL-3 genomic libraries.	72
Chapter 3	<u>Molecular structure of a human complement component C4A gene</u>	
3.1	Introduction	74
3.2	Subcloning, restriction mapping and DNA sequencing	77
3.3	Results and discussion	78

3.3.1	The 5' regulatory region	78
3.3.2	The exon/intron structure of the C4A gene	80
3.3.2.1	Structure of the β chain	80
3.3.2.2	Structure of the α chain	82
3.3.2.3	Structure of the γ chain	89
3.3.3	The 3' end of the C4 genes	90
3.3.4	Exon phases and the amino acids encoded at the exon-intron splice junctions	91
3.3.5	Comparison of the exon-intron structure among human C4, mouse C4 and mouse C3 genes at the 5' regions	92
3.4	Summary and conclusion	92
Chapter 4	<u>Structural basis of the polymorphism of human C4A and C4B: gene size, reactivity, and antigenicity</u>	
4.1	Introduction	96
4.2	Results	99
4.2.1	Cloning, mapping and assignment of C4 genes	99
4.2.2	Sequence analysis of 9 allotypes of C4	102
4.2.2.1	C4A 3a	104
4.2.2.2	C4B 3	105
4.2.2.3	C4A 1	106
4.2.2.4	C4B 5	107
4.2.2.5	Comparison of intron sequences of C4d	108
4.3	Discussion	108
4.4	Summary	116

Chapter 5	<u>A structural model for the location of the Rodgers and the Chido antigenic determinants and their correlation with the human C4A/C4B isotypes</u>	
5.1	The Rodgers and the Chido antigens are human C4 antigens	117
5.2	Polyspecificity of the alloantisera and multiple determinants for Rg and Ch	118
5.3	A structural model for the C4 isotypic and antigenic determinants	121
5.4	Prospectives and implications	125
5.5	Summary	128
Chapter 6	<u>Definitive RFLPs to distinguish between the human C4A/C4B isotypes and the major Rodgers/Chido Determinants: application to the study of the C4 null alleles</u>	
6.1	Introduction	129
6.2	Results	132
6.2.1	C4A/C4B isotypicity- Nla IV polymorphism	132
6.2.2	Rg1 and Ch1 antigenicity- Eco O109 polymorphism	133
6.2.3	Applications: studies on the C4 null alleles	134
6.3	Discussion	136
6.3.1	RFLPs of human C4	136
6.3.2	C4 Q0 alleles	138
6.3.3	The two loci of C4	139
6.4	Summary	140
Chapter 7	<u>General Conclusions</u>	142

REFERENCES

APPENDIX I Genomic DNA and derived amino acid sequences of
the C4A 3 gene from Cos 3A3

APPENDIX II Publications

CHAPTER ONE

INTRODUCTION

1.1 General introduction

The structural genes for the human complement components C4, C2 and Factor B are located in the class III region of the human major histocompatibility complex (MHC). These class III gene products exhibit various degrees of polymorphism and that of C4 is the most complex and fascinating.

There are two isotypes of C4, C4A and C4B, which have different haemolytic activities. They are related to the plasma and the erythrocyte antigens: the Rodgers (Rg) antigen to C4A and the Chido (Ch) antigen to C4B. Based on the differences of the electrophoretic mobilities due to variations in charge and molecular weight (mol. wt.), ~10 to 20 allotypes have been detected in each isotype. Some of the allotypes have lower reactivities as compared with others of the same isotypes. Furthermore some rare variants express the reversed association with the Rg/Ch antigens. One puzzling phenomenon in the studies of the genetics of the ^{Human leukocyte antigen,} HLA is the increased frequency of some specific haplotypes to many hereditary and/or autoimmune diseases. Some HLA genes are in linkage disequilibrium. The extensive polymorphic pattern of human C4A/C4B can be used as a useful genetic marker to follow the inheritance of these diseases, and may also help to trace the origin of the HLA-related diseases, in parallel with the studies on class I and class II products. Whether the polymorphism of C4 and those of the HLA class I and class II products are interrelated in the network of immune response is

unknown.

Another interesting aspect of the C4 molecule is its covalent binding ability to nearby molecules, after activation, through the formation of amide or ester bonds. This is due to the presence of a labile thiolester bond, which is also present in complement component C3, and the protease inhibitor α_2 macroglobulin. During the activation of C4, a small peptide of 77 amino acid residues, C4a, is released. This peptide is an anaphylatoxin. A similar event is also observed during the activation of C3 and C5, but not in α_2 macroglobulin. Structural studies have suggested that C4, C3, C5 and α_2 macroglobulin are evolutionarily related, although they have different biological roles, and their corresponding genes are located on different chromosomes in man.

In 1983, the human complement C4 genes were cloned (Carroll and Porter, 1983). The DNA sequence of a C4 cDNA and the precise locations of the C4 genes with respect to the C2 and factor B genes were elucidated in the subsequent year (Belt et al., 1984; Carroll et al., 1984a). This opened up the study of the genetics and polymorphism of C4 at the molecular level.

In 1977, it became clear that eukaryotic genes are split by intervening non-coding sequences or introns. Gilbert (1978, 1985), Go (1981) and Blake (1979) proposed that the coding sequences in pieces, i.e. exons, may reflect that proteins are also in pieces. In other words, exons may be related to domains, sub-domains, modules or secondary folding of a protein. New proteins may be evolved from a combination, or

shuffling of exons from other existing proteins. Thus, determination of the exon/intron structure of a gene may help to trace the evolution of a protein and its relationship with other proteins. Together with the 3-dimensional protein structure, it may help to correlate the structure and function of a protein with greater confidence.

The genomic sequence of a human C4A gene was determined in this work. The primary objective was to elucidate the exon/intron structure of the C4A gene, with reference to the cDNA sequence. It was hoped that this would also provide background information for further studies on the polymorphism of human C4: to elucidate the molecular basis of the isotypic, antigenic and allelic differences and for the exploration of the origin of some HLA-linked genetic diseases.

1.2 The complement system

1.2.1 Introduction

The complement system is the principal effector mechanism of vertebrates in the immune defence against infection by microorganisms. Activation products of the complement components cause lysis of cellular antigens, attract phagocytic cells to the site of activation, facilitate uptake and destruction of the complement activators by the phagocytes, and solubilize immune complexes by regulating their size and promote clearance via the liver. The complement system in human plasma is activated in response to the formation of antibody-antigen complexes, or the entry of various types of foreign material into the circulation or

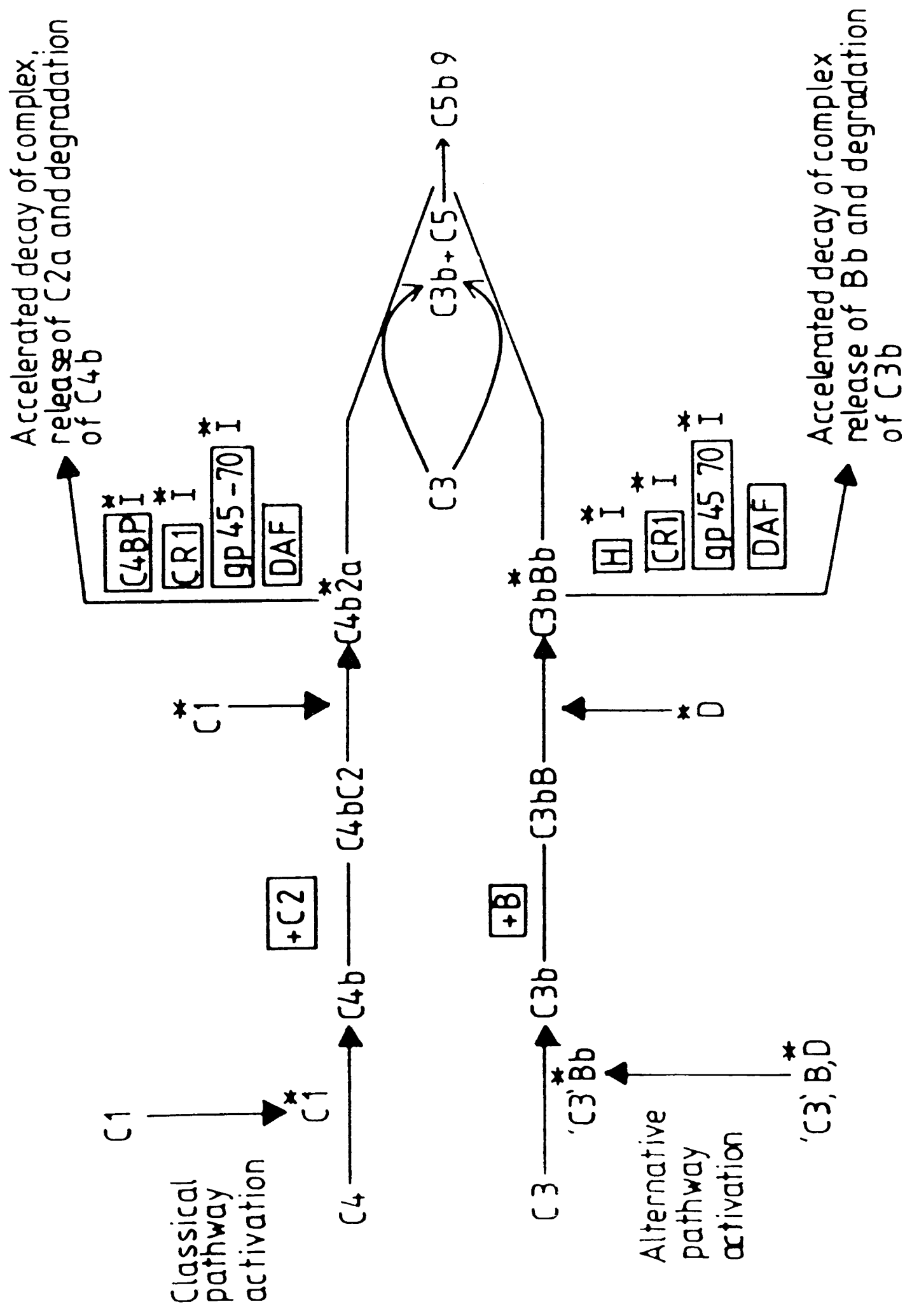
tissue spaces. Various aspects of the human complement system have been reviewed by Reid and Porter (1981); Reid (1983, 1986); Porter (1983, 1984, 1985); Campbell et al., (1986), Lachmann (1979, 1984); Lachmann and Hugh-Jones (1984); Lachmann and Hobart (1985); Cooper (1985); Muller-Eberhard (1986); Fearon and Wong (1983); Roos and Medof (1985); Sim et al., (1986).

There are two pathways of activation, i.e. the classical and the alternative pathways, involving more than 13 components. A schematic presentation of the complement activation pathways is shown in Fig. 1.1. In both cases the common feature is a sequence of proteolytic cleavages of zymogens to active components, resulting in the assembly of a metastable, multi-component C3 convertase, which has equivalent specificity but different compositions. C3 convertase activates C3, which binds to the enzyme complex and changes its specificity to that of C5 cleavage. The classical pathway C5 convertase is C4b2a3b and that of the alternative pathway is C3bBbC3b. In both C3 and C5 convertases, the enzymic components lie in the homologous protein fragments, C2a and Bb. The convertases are anchored to the surface of the target cells via covalent ester or amide bondings from C3b and C4b. Upon activation of C5, five proteins, C5b, C6, C7, C8 and C9, interact in a sequential manner and form the macromolecular organisation known as the membrane attack complex (MAC). MAC will form a transmembrane channel and lyse the target cell.

Activation of the two complement pathways can be achieved

Fig. 1.1 Main steps in the activation of the classical and
alternative pathways of complement

Proteins which interact with C3b, or C4b, are enclosed in boxes. Enzymically active components are denoted by asterisk.
(After Reid et al., 1986)



in an antibody dependent and an antibody-independent manner. The classical pathway is generally triggered by antibody-antigen complex. The components involved include the C1 complex (i.e. C1q, C1r and C1s), C2 and C4. The alternative pathway can be initiated secondarily by the classical pathway, but it can also be triggered independent of antibody by foreign polysaccharides such as those found on bacterial and yeast cell walls. It is thought that the alternative pathway of complement activation forms the first line of defence against infection prior to an immune response. The complement components involved in the alternative pathway are factor D, factor B and C3b. A positive feedback amplification loop is present in this pathway so that the activation process can be very rapid and efficient. It results in attachment of a large number of C3b to the activator surface. This surface bound C3b reacts with a variety of cells bearing specific receptors and thus promote the adherence phase of phagocytosis.

The precise control of both activation pathways is very important as excessive activation may cause tissue damage, and inefficient activation will slow the dissolution and removal of immune aggregates, resulting in lupus diseases. The fine control mechanism is mediated by seven humoral and many membrane-bound regulatory proteins or receptors at various stages of activation.

Inherited deficiencies in most of the complement proteins have been reported. Individuals deficient in complement proteins suffer a variety of diseases with different severities. Patients with inherited deficiency in C3 are

subject to recurring bacterial infection. This emphasises again the importance of the complement system.

1.2.2 The classical pathway

1.2.2.1 The C1 complex

Activation

The classical pathway of complement activation is triggered by the binding of C1 to immune complexes, or aggregates containing IgG1, IgG2, IgG3 and IgM. It can also be activated independently of antibodies by a variety of polyanions, small polysaccharides, retroviruses, mitochondrial membranes, etc. (reviewed by Cooper, 1985). The activation mechanism by antibody-antigen complexes is well-studied and is briefly described as follows.

C1 is a multi-molecular complex consisting of two loosely interacting entities; C1q, the recognition subunit, and the catalytic subunit which is a calcium-dependent, tetrameric association of two serine proenzymes, C1r and C1s. Binding of C1 to various activators is mediated by C1q which serves as an allosteric effector for the activation of C1r and C1s. Activation of the catalytic subunit is a two-step process: (1) autolytic activation of proenzyme C1r to its active form C1r; and (2) conversion of proenzyme C1s to C1s by C1r. C1s will subsequently activate C4 and C2 to form the classical pathway C3 convertase. Thus C1r works as a bridge in transferring the activation signal (i.e. the fixation of C1q on activating agents) into an enzymic activity (on C1s). Binding of C1q to IgG is through the globular heads of C1q to the C_H2 domain of

IgG, presumably through ionic interactions. Divalent binding to the C1q molecule by two IgG molecules is necessary to activate the C1 complex.

C1q

The C1q molecule is made up of three types of similar polypeptides, A, B, and C, each about 225 amino acid residues in size. C1q is composed of 18 polypeptide chains with a composition of 6A, 6B and 6C. These peptide chains forms 6 A-B and 3 C-C dimers through disulphide bridges. Based on data from biochemical and electronmicroscopic analysis, Reid and Porter (1976) proposed a structural model for the C1q molecule (Fig.1.2). The A, B, and C chains together form a collagen-like structure at the N-terminal portion and a globular head at the C-terminal regions. Two of these ABC triplets are connected together to form a structural subunit by disulphide linkages through the two C chains. Three of these structural subunits assemble through non-covalent interactions to form the 'tulip-like' structure of C1q. The collagen-like triple-helix is due to the presence of the collagen-specific, repeating amino acid sequences (i.e. Gly-X-Y) at the N-terminal regions of the A, B and C chains. About halfway along the length of each collagen region, the repeating units are interrupted and thus the triple helical structure may be disrupted. This disruption corresponds to the place where the helical structure bends to form the 'tulip-like' structure. Interestingly, this position also corresponds to the exon-intron splice junction of the C1q B-chain gene, which contains only a single intron (Reid, 1985).

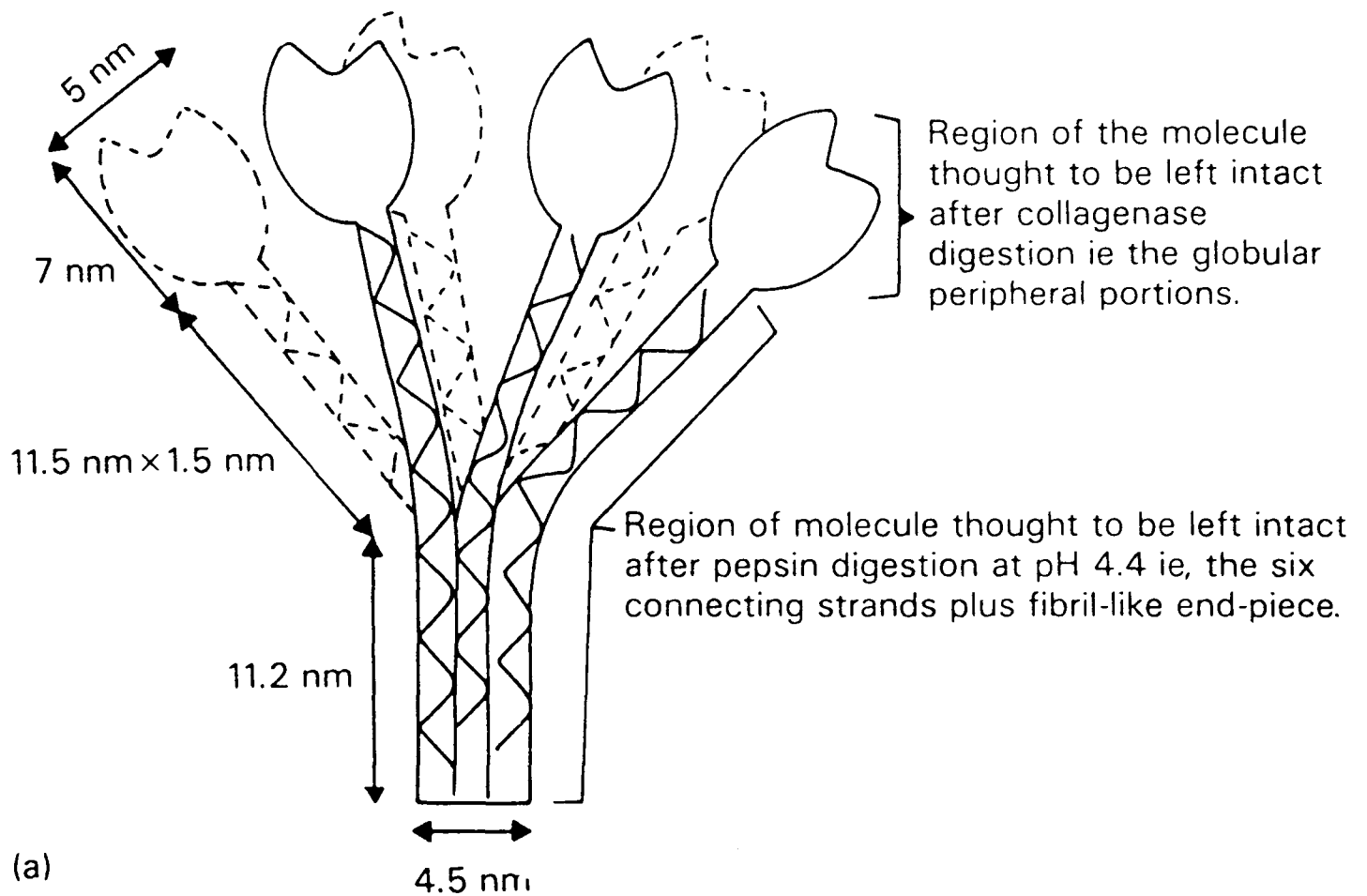
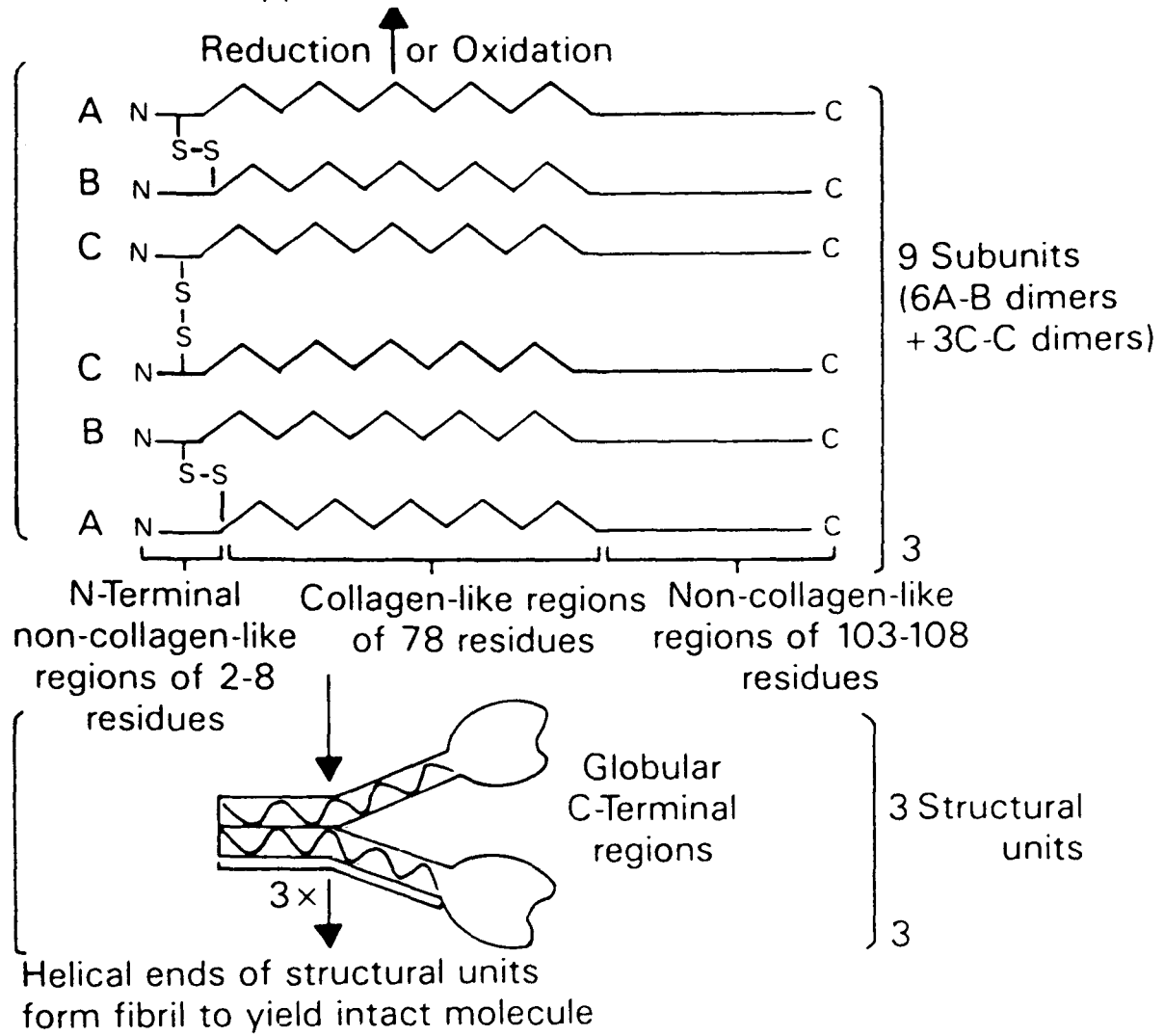
Fig 1.2 Structure of human C1q

The wavy line represents the proposed triple helix section, i.e. collagen-like and fibril-like end-piece. Solid line, towards the reader; dotted line, away from the reader. (After Reid, 1986).

STRUCTURE OF HUMAN C1q

6A + 6B + 6C Chains
(each approx. 23000 mol. wt.)

18 chains



C1r and C1s

C1r and C1s are homologous molecules because they share about 45% identity in amino acid sequence (Carter et al., 1984) and similar domain structures viewed under the EM. Both C1r and C1s are serine proteases. They interact in a calcium dependent manner to form the tetrameric structure, C1s-C1r-C1r-C1s. On activation, the two non-covalently bound C1r zymogen polypeptide chains have intrinsic and specific proteolytic activity for a peptide bond on the adjacent subunit (C1r), thus forming the disulphide-linked A and B chains. Enzymic activity resides on the smaller B chain. C1s is the only physiological substrate of C1r. C2 and C4 are the only substrates of C1s. A functional model for the structure of the C1r and C1s tetramer and their interaction with C1q has been proposed by Colomb and coworkers (Colomb et al., 1984; Arlaud et al., 1987a).

The complete amino acid sequence of C1r has been established by protein sequencing (Arlaud and Gagnon, 1983; Arlaud et al., 1987b) and confirmed by derived cDNA sequences (Journet and Tosi, 1986). It was found that in the C1r A chain, there are two internal repeating segments that share a 30% identity in about 70 residues. It was also found that there is an epidermal growth factor conserved sequence, and two tandemly located sequence elements that is characteristic of C3b/C4b binding proteins.

Control of the C1 activation

Activation of C1 is finely regulated by at least 3 mechanisms: (1) intrinsic lability of the C1 complex, (2)

negative control by the C1 inhibitor and (3) negative feedback regulation by nascent C3b and C4b molecules (reviewed in Cooper, 1985). C1 inhibitor is a member of the serine protease inhibitor family (Tosi et al., 1986). It binds rapidly and irreversibly to activated forms of C1r and C1s, thereby inhibiting their complement function. Its binding also enhances the dissociation of C1r and C1s from the C1 complex (in the form of C1In-C1r-C1s-C1In). It has been found that nascent C4b specifically inhibits the turnover of the IgM-activated C1 complex, while C3b inhibited the turnover of the IgG-activated C1 complex. This appears to be a negative feedback mechanism to circumvent uncontrolled activation of C1, as C4b and C3b are products of the activation pathway (Ziccardi, 1986).

1.2.2.2 The assembly of the C3 convertases and C5 convertases

The classical pathway C3 convertase is a Mg^{2+} -dependent complex of C4b and C2a generated by limited proteolysis of C4 and C2 by activated C1. C4b anchors on the membranes, while C2a bears the active serine protease site of the enzymatic complex.

C4 is synthesised as a single-chain precursor of mol. wt. 200,000 and is post-translationally processed to a three-chained structure linked by disulphide bridges. The protein contains about 7% carbohydrate, and the order of the three chains is β - α - γ , which have mol. wt. of 93,000, 75,000 and 33,000, respectively (Schreiber and Muller-Eberhard, 1974;

Gigli et al., 1977; Hall and Colten, 1977; Chan and Atkinson, 1983; 1985). Human C4 is activated by $\overline{\text{C1s}}$ which cleaves the chain to release the activation peptide C4a. The major moiety, C4b, can bind covalently to a variety of surfaces via a reactive glutamyl group from the thiolester in its α' chain (see section 1.4 for further description).

C2 is a single chain glycoprotein of mol. wt. 100,000. Protein sequencing and derived amino acid sequence from cDNA clones showed that the zymogen form of C2 is 732 amino acids long (Christie et al., 1980; Gagnon, 1984; Bentley and Porter, 1984; Bentley, 1986). In association with C4b and Mg^{2+} , C2 is cleaved by $\overline{\text{C1s}}$ to form the catalytic chain C2a that contains the serine protease domain (509 residues), and the N-terminal peptide C2b. It was suggested that C2 binds to C4b non-covalently at least at two sites, one from C2b and one from C2a (Nagasawa and Stroud, 1977; Kerr, 1980). Transmission electronmicroscopic studies revealed that the C2 molecule is composed of three globular domains, one from C2b and two from C2a. Each of these globular domains is about $40\text{\AA}$ in size. It was also found that C2a only associates with C4b through one of the two domains. The other 'free' domain is assumed to contain the active site of the serine protease. C4b appears as an irregular, multiple-domain structure in the EM (Smith et al., 1984a). The role of Mg^{2+} in the assembly and function of the C3 convertase is still uncertain.

The C4b2a convertase is highly unstable with a half-life of only a few minutes (Polley and Muller-Eberhard, 1967; Kerr, 1980). The instability is due to the decay-dissociation of

C2a from the complex. Once dissociated, C2a cannot reassociate with C4b to form an active enzyme. Besides this temperature dependent intrinsic decay, the complex enzyme is subjected to extrinsic decay by the plasma C4b binding protein (C4bp). The binding of C4bp to the C3 convertase also enables the proteolysis of C4b at two sites of the α' chain by the C3b/C4b inactivator, factor I. In the latter reaction C4bp works as a cofactor of factor I, which is also a serine protease. The cleavage reaction releases the 45,000 mol. wt. C4d fragment from the major moiety, C4c (Fujita and Nussenzweig, 1979; Gigli et al., 1979; Cooper, 1975). Thus C4bp and factor I are important humoral homeostatic controlling agents of the classical pathway C3 convertase.

Membrane bound C3 convertase is controlled by (1) CR1 which, besides its receptor function, also accelerates decay-dissociation of C2a and possesses cofactor activity for factor I; (2) decay accelerating factor (DAF), which accelerates decay of the enzyme complex and protects the body from cell damage by autologous, activated complement factors and (3) membrane cofactor protein, MCP (previously called gp45-70), which serves as ^a₁ cofactor for factor I to inactivate C4b (reviewed by Holers et al., 1985).

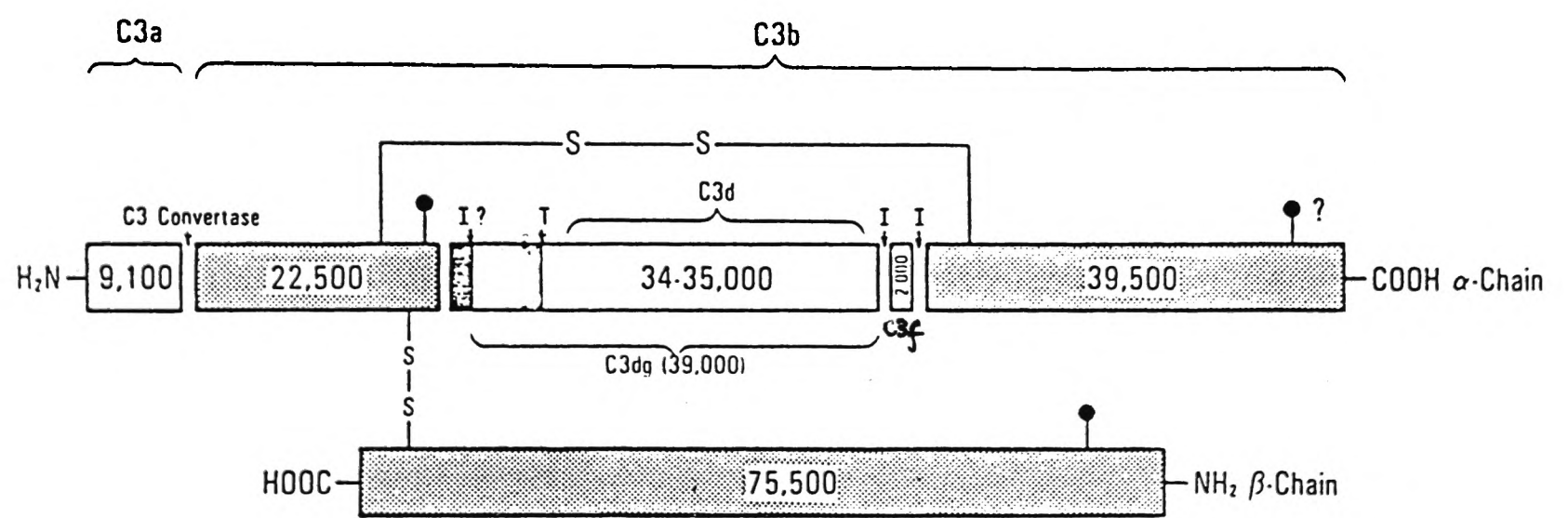
The relatively low titer of plasma C2, the complexity and the short half-life of the C3 convertase altogether imposed considerable difficulty on the biochemical studies of the enzyme complex. This was in part circumvented by (1) mild oxidisation of C2 by dilute iodine solution (i.e. C2ox) prior to its assembly with C4b (Polley and Muller-Eberhard, 1967);

(2) substitution of Mg^{2+} by Ni^{2+} in the reaction buffer for the convertase assembly (Fishelson and Muller-Eberhard, 1982) and (3) addition of the convertase with the classical pathway nephritic factor which is a IgG auto-antibody isolated from a patient suffering post-streptococcal glomerulonephritis (Halbwachs et al., 1980; Gigli et al., 1985). Stabilisation of the C3 convertase by iodine was suggested due to the formation of a covalent bond on C2a presumably on a site which may be related to the decay-dissociation of the enzyme complex (Parkes et al., 1983). The role of Ni^{2+} on stabilisation of the C3 convertase is unknown. It was found that the effect of C2ox and Ni^{2+} substitution on the half-life of the convertase is additive (Villiers et al., 1985). The classical pathway nephritic factor can stabilise C3 convertase by preventing its intrinsic decay, and inhibiting the extrinsic decay by C4bp and proteolysis by factor I/C4bp (Gigli et al., 1985).

C3 is the most abundant complement protein in the plasma, ~1.3 g/l. It is synthesised as a single chain precursor molecule and is processed to the disulphide linked β - α chains before secretion. The complete cDNA sequence and its derived amino acid sequence of human C3 have been determined (DeBruijn and Fey, 1985). A schematic representation of the protein structure of C3 is shown in Fig. 1.3. The precursor C3 consists of 1641 amino acid residues. The β chain (645 residues) is located at the N terminal region and the α chain (992 residues) is located at the carboxyl end. C3 plays a central role in the complement system since its proteolytic

Fig. 1.3 Schematic representation of the mature human C3 (after deBruijn and Fey, 1985)

Cleavage of C3 by C3 convertase liberates the anaphylatoxin C3a from the N-terminus of the α chain and generates C3b. Cleavage sites for factor I (I) and trypsin (T) define the sub-fragments C3dg, C3d and C3f as indicated. The shaded area represents C3c. The precise locations of the interchain disulphide linkages are unknown. Glycosylation sites are marked by $\bullet$. Molecular sizes (in daltons) for various fragments were deduced from amino acid sequence and do not include carbohydrate content. The 2000 mol. wt. fragment formed by factor I cleavage (with factor H, or CR1, or MCP as cofactor) is called C3f. A further cleavage site, I(?), is mediated by factor I with CR1 as cofactor (Medof et al., 1982).



cleavage is an event common to both activation pathways, and many of the biological properties of complement are mediated by C3 and its cleavage fragments (Muller-Eberhard, 1981).

Cleavage of C3 by the C3 convertases give rise to two activated fragments, the anaphylatoxin, C3a which is a vasoactive peptide and a mediator of inflammation, and C3b which can act as a cofactor in the formation of the C5 convertases. C3 like C4 has an intra-chain thiolester bond (Law & Levine 1977, 1980b; Tack et al., 1981). The binding of one or more activated C3 molecules to, or adjacent to, the C3 convertase alters its specificity to a C5 convertase (Cooper and Muller-Eberhard, 1970). The active site of the C5 convertase, C4b2a3b, is in the C2a fragment, and the bound C3b serves as a binding site for C5 (Vogt et al., 1978). The nature of the C5-binding site in the C5 convertase has been examined recently by Takata and coworkers (1987). It was found that during the activation of C3, one of the nascent C3b molecules generated by the C3 convertase directly binds covalently to C4b and that this C4b-C3b dimer serves as a high affinity binding site for C5. Binding of C5 to the C3 convertase is probably divalent in that C5 recognises both protomers in the C4b-C3b dimer. This enables C5 to bind selectively to the complex convertase instead of the surrounding monomeric C3b. Thus the activation of C5 and the membrane attack complex formation is precisely located at the site of the complement activation.

The covalent attachment of C3 and C4 to foreign particles also causes partial dissolution (Takata et al., 1984) and

speeds their removal due to the presence of C3 and C4 receptors on the surfaces of erythrocytes and phagocytic cells (reviewed in Fearon and Wong, 1983; Roos and Medof, 1985).

1.2.2.3 The regulatory proteins of the classical pathway C3 convertase

As already mentioned the classical pathway C3 convertase is controlled by at least two plasma proteins (ie. factor I and C4bp) and three membrane-bound proteins (i.e. CR1, DAF and MCP) to prevent the deleterious effect caused by over activation of the complement system and the deposition of activated components on autologous tissues.

Factor I

Factor I is a member of the serine proteases. In the presence of C4bp, or CR1, (or MCP), factor I inactivates C4b by two proteolytic cleavages on the α' chain to produce C4d and C4c. Unlike many serine proteases, it does not appear to have a circulating proenzyme form and it is not inhibited by any of the known plasma proteinase inhibitors (Crossley, 1981). Full length cDNA clones for factor I have been obtained (Catterall et al., 1987). Sequence determination reveals that the light chain (224 residues) has clear homology with other serine proteases. The heavy chain (321 residues) contains the cysteine-rich, low-density lipoprotein (LDL) class A and class B repeats, which is also present in the C1r A chain, complement C9, C8 α and C8 β (Stanley et al., 1985; Howard et al., 1987; Rao et al., 1987).

An interesting aspect for the C3b/C4b binding proteins

Table 1.1 Proteins for which there is structural evidence for repeating homology units

Protein		Interacts with							
<i>Complement Proteins</i>									
Factor B	Three contiguous units starting at N-terminus of 92 000 M, chain	C3b							
C2	Three contiguous units starting at N-terminus of 108 000 M, chain	C4b							
C4b-binding protein (composed of 7 identical chains)	Eight contiguous units starting at N-terminus of 70 000 M, chain	C4b							
Factor H (mouse)	Twenty contiguous units throughout entire 160,000 M, chain	C3b							
C3b:C4b receptor (CR1)	Number not yet established but, at least eight units or perhaps up to thirty-six throughout 160–250 000 M, chain	C3b/C4b							
Subcomponent C1r	Two units located near C-terminus of non-catalytic 56 000 M, chain	Not known							
<i>Non-Complement Proteins</i>									
β ₂ glycoprotein I	Five contiguous units starting at N-terminus of 50 000 M, chain	Not known							
IL-2 receptor	Two units starting at N-terminus of 55 000 M, chain (the two repeating units are separated by domain of 37 amino acids)	Not known							
β subunit of clotting factor XIII	Ten contiguous units starting at N-terminus of 80 000 M, chain	Not known							
General consensus sequence of approximately 60 amino acids* seen in most repeating homology units:									
4	7	30	32	35	46	50	52	57	59
CYS	PRO	TYR PHE	CYS	GLY	CYS	GLY	TRP	ALA PRO	CYS

such as C4bp, CR1, DAF and factor H (which is an analogous protein of C4bp in the alternative pathway) is that they are tightly linked on the long arm of human chromosome 1 (region q31-32) (Rodriguez de Cordoba et al., 1987; Weis et al., 1987; Lublin et al., 1987) and that they all possess the ~60 amino acid repeat which is also present in factor B, C2, C1r, C1s and some non-complement proteins as shown in Table 1.1 (reviewed by Reid et al., 1986). This repeating unit is characterised by a consensus sequence having a framework of highly conserved residues consisting of one tryptophan, two proline and four cysteine residues. It also appears that each of these repeating unit is encoded by a separate exon (Campbell et al., 1984; Lintin and Reid, 1986). The functional role of the ~60 amino acid residues repeat is unknown but for the case of the complement proteins, they may be related to the formation of the C3b/C4b binding domain(s) or they may correspond to a specific structural module (Reid et al., 1987) similar to that of the immunoglobulin fold (Williams, 1984). For convenience, this repeating unit is designated as the short consensus repeat (SCR) in this thesis.

C4bp

C4bp has a mol. wt. of about 550,000 and is formed by 7-10 identical peptide chains. Derived amino acid sequence from the cDNA together with amino acid sequencing of human C4bp (Chung et al., 1985a,b) have given the complete amino acid sequence of 549 residues. The N terminal 491 amino acids are comprised of 8 repeating units previously described. Electron microscopic studies revealed that C4bp has a 'spider-

like' (Dahlback et al., 1983) or 'bunch of flowers' (Villiers et al., 1985) structure. In other words, it is a polymeric structure with a central core structure and many flexible (7-10) tentacles. These tentacle-like elements may be formed by the N-terminal repeating units of each monomer which are probably involved in the C4b binding. The central core is probably formed by the C-terminal regions of each monomeric subunit linked by disulphide bonds (Villiers et al., 1985b; Reid et al., 1986). It was suggested that each C4bp complex can bind 4-6 C4b molecules in the fluid phase conditions (Ziccardi et al., 1984).

CR1

The human C3b/C4b receptor (CR1) was firstly described at the molecular level by Fearon (1979). It is an integral membrane glycoprotein that is expressed on cells of immunological (T and B lymphocytes, monocytes, mast cells, polymorphonuclear cells) and nonimmunological (primate erythrocytes, glomerular podocytes) lineages. Functions mediated by CR1 include phagocytosis of erythrocytes and bacteria, endocytosis of small immune complexes, and generation of second signals. CR1 binds C4b and C3b that have covalently attached to immune complexes and other complement activators, and the consequences of these interactions depend upon the cell-type bearing the receptor (reviewed by Fearon and Wong, 1983; Fearon, 1984; Ross and Medof, 1985; Sim et al., 1987). Erythrocyte CR1 binds complement component C3b- or C4b- bearing immune complexes and may transport them to the mononuclear phagocytic system in the liver and spleen

(Cornacoff et al., 1983). As already mentioned, CR1 is a cofactor for factor I mediated cleavage of C4b (and C3b in the alternative pathway), and accelerates the decay of the classical (and the alternative) pathway C3 convertase. The concerted activity of CR1 and factor I in the degradation of C4b is very effective in the endogenous environment (i.e. when both CR1 and the deposited C4b are on the same cell membrane). Thus CR1 plays an important role to inactivate the complement cascade on autologous host membrane (Kinoshita et al., 1986).

Derived amino acid sequences representing ~80% of the most common allotype of CR1 reveals a striking feature on the primary structure of CR1 (Klickstein et al., 1987). It was found that CR1 is made up of at least 30 short consensus repeating units (SCR) in addition to the transmembrane and cytoplasmic domain sequences. The SCRs are further organised to four higher order long homologous repeating units (LHRs). Each LHR is made up of seven SCRs and is ~450 amino acid in size. Amino acid identity between LHRs ranges from 70% between the first and third repeats to 99% between the N-terminal 250 amino acids of the first and second repeats. Two additional SCRs join the four LHRs to a single membrane-spanning domain of 25 amino acids, which is linked to the 43 residues carboxyl terminal cytoplasmic domain that is partially homologous to epidermal growth receptor.

DAF and MCP

DAF is a 70,000 dalton glycoprotein which appears in two forms, ie. the membrane-bound and the soluble, secretory form (Medof et al., 1987). The membrane-bound DAF is present on

all cell membranes in contact with complement. It is anchored to the cell membrane by phosphatidylinositol (Medof et al., 1986; Tse et al., 1985). Interaction of DAF with autologously cell-associated C4b (and C3b) polypeptides inhibits the uptake of C2 (and factor B) and thus prevents the assembly of the C3 convertase(s). It halts the progression of the complement cascade and prevents the consequent, autologous cell-injury, without affecting the structure of C4b (and C3b) (Klinoshita et al., 1986). In other words, DAF acts only on surface of the same cells, it does not possess cofactor activity with factor I.

Complementary DNA clones of human DAF have recently been obtained (Caras et al., 1987; Medof et al., 1987). Derived amino acid sequence suggests that there are four contiguous SCR near the N-terminal region of the protein. Two types of cDNA clones have been obtained, which are suggested to be the result of differential RNA splicing (Caras et al., 1987). The shorter form may correspond to the membrane bound DAF (347 residues) as it contains a hydrophobic C-terminal region. The larger form contains an unspliced 'intron' with the Alu-repetitive sequence corresponding to the C-terminal region of the protein. This unspliced intron may drastically decrease the hydrophobicity of the protein near the C-terminus and may correspond to the soluble form of DAF (406 residues). There are clusters of serine and threonine residues in both forms of DAF, which may correspond to the O-linked glycosylation sites.

MCP was previously called gp 45-70 to indicate its mol. wt. on SDS-PAGE and later renamed membrane cofactor protein

because of its functional activity (Seya et al., 1986 a) This protein is present on human platelets, monocytes, B and T lymphocytes but not on erythrocytes. MCP appears to function only as a surface-bound cofactor for the first cleavage of factor I-mediated breakdown of C4b and C3b. Thus it is complementary to DAF in function. Compared with CR1, it does not have decay acceleration activity and it does not mediate the adherence of C3b- or C4b- coated particles. It seems that there are two allotypes: one with mol. wt. of 63,000 and the other with mol. wt. 58,000, which assumes an autosomal codominal manner of inheritance (Ballard et al., 1987). The structural gene for MCP has not been cloned, nor its chromosomal location been mapped. It will be of interest to determine whether this protein also maps to the C3b/C4b regulatory protein complex on the long arm of chromosome 1.

It seems that the functions among the three membrane-bound regulatory proteins, CR1, DAF and MCP overlap. However DAF and MCP may be mainly responsible for the intrinsic decay-accelerating activity, or factor I cofactor activity, while CR1 may be mainly responsible for the extrinsic decay and cofactor activity in addition to its paramount role in the immuno-clearance mediated by the receptor activity (Ballard et al., 1987). These membrane-bound proteins appear to be very important to the protection of the host body against cell damage caused by autologous complement activation, or against aggregated immune complexes. Patients deficient in CR1 result in systemic lupus erythematosus (SLE) (Dykman et al., 1984) while patients deficient in DAF may suffer paroxysmal

nocturnal haemoglobinuria (Medof et al., 1985).

1.2.3 The alternative pathway of activation

The alternative pathway can be activated by a large number of very different activators such as particulate polysaccharides, bacterial and yeast cell walls, parasites, virus infected mammalian cells, aggregated antibodies, etc. As a result, it is difficult to generalise the common features among these activators except that they are mostly polymeric with some types of repeating structures, and are particulate in nature (reviewed in Reid and Porter, 1981; Lachmann and Hugh-Jones, 1984). The components involved in the alternative pathway include factor D, factor B and C3.

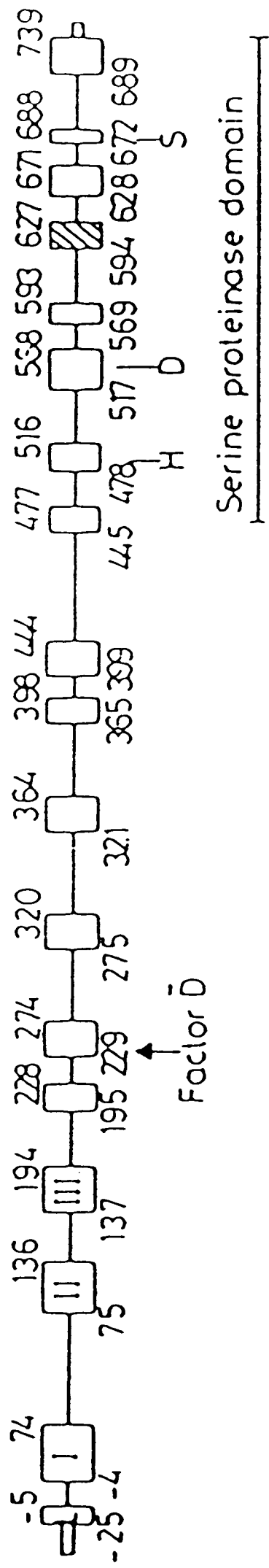
Factor D is a serine protease with ~25,000 mol. wt. present in plasma in activated form. Factor D catalyses the conversion of factor B complexed with C3b and Mg^{2+} to form the alternative pathway C3 convertase, C3bBb. This can cleave C3 via the active site in Bb to form C3a and C3b. C3b can associate with C3bBb to form the alternative pathway C5 convertase, or it can assemble with factor B to amplify the amount of C3 convertase near the site of activation. The structural gene for factor D is yet to be cloned. Factor B is a serum glycoprotein with mol. wt. ~90,000, containing ~7% carbohydrate. EM studies showed that factor B is similar to C2 in morphological structure in having a three-globular-domain structure, i.e. one corresponds to Ba and two correspond to Bb (Smith et al., 1984). Upon binding to C3b, factor B is rendered susceptible to the proteolytic cleavage

by factor D, forming a 30,000 mol. wt. Ba fragment and a 60,000 mol. wt. Bb fragment. The serine protease domain resides in the Bb subunit (reviewed in Gagnon, 1984). Attachment of factor B to C3b is mediated by two points of contact, one being in the Ba domain, and the other in the Bb domain (Prydzial and Isenman, 1987). Human cDNA and genomic clones of factor B have been obtained (Woods et al., 1982; Campbell and Porter, 1983; Campbell et al., 1984; Morley and Campbell, 1984). Derived amino acid sequence suggested that factor B consists of 739 residues and is encoded by 18 exons of the factor B gene (Fig. 1.4). The Ba domain is encoded by exons 2-6, of which three correspond to the SCR (Morley and Campbell, 1984). These three repeating units are encoded by exons 2,3 and 4, respectively. The serine protease domain is encoded by exons 11-18. The three amino acid residues essential for catalysis, i.e, His 501 Asp 551 and Ser 674 are encoded by separate exons (Campbell and Porter, 1983).

The fluid phase and cell-bound alternative pathway C3/C5 convertases are very unstable complexes and are subjected to intrinsic decay by the dissociation of Bb. They can be extrinsically decay-dissociated and/or proteolytically inactivated by factor H and factor I in plasma, or DAF, MCP, CR1 and probably CR2 together with factor I on cell membrane. They can be stabilised by properdin. Properdin functions by binding to and stabilising the alternative pathway C3/C5 convertases. There is no analogue of properdin in the classical pathway. In plasma, properdin appears as a mixture of polymeric glycoprotein chains each of 56000 mol. wt. Monomeric properdin appears as a highly asymmetric structure

Fig. 1.4 Structure of the factor B gene

Exons are shown boxed; the number refers to the amino acids encoded by each exon. L denotes the putative leader peptide. The exons encoding the three homologous regions (i.e. SCR) in the Ba domain are labelled I, II, and III. The serine protease domain is encoded by the last eight exons. H, D, and S denote the positions of the codons for the active site residues His 501, Asp 551 and Ser 674, respectively. The shaded exon is not found in many serine proteases such as trypsin and chymotrypsin. The factor D cleavage site which separates Ba and Bb coding regions, is marked with an arrow. (After Campbell and Bentley, 1985).



1 kb

in the EM (Smith et al., 1984b).

Factor H is analogous to C4bp in the classical pathway as it accelerates decay-dissociation of Bb from C3bBb and also acts as a cofactor for factor I mediated proteolytic cleavage of C3b to form iC3b. The latter process releases a small peptide ~2,000 mol. wt., C3f. Derived amino acid sequences from full-length murine cDNA and partial human cDNA clones of factor H have revealed that the protein is entirely composed of 20 repeating units characteristic for C3b/C4b binding proteins (Kristensen and Tack, 1986; Ripoché et al., 1986). The binding site for C3b was mapped to a 38,000 mol. wt. tryptic fragment (Alsenz et al., 1985) which is probably encoded by repeats I to V.

Membrane-bound DAF decays the alternative pathway C3 convertase and enhances the dissociation of the enzymic Bb subunit, (as in the case of C2a in the classical pathway convertase). MCP is complementary to DAF as it serves as a cofactor for the factor I mediated cleavage of C3b to iC3b, with an activity ~50 times more potent than factor H in terms of quantity (Seya et al., 1986a). CR1 possesses both decay-dissociation activity and factor I-cofactor activity for the convertases. The inactivation of C3b by CR1 is unique in two ways. Firstly, it enables the factor I mediated cleavages of C3b to iC3b, and iC3b to C3c and C3dg (Medof et al., 1982). Secondly, CR1-factor I mediated C3b cleavages also works extrinsically, i.e. on other cells deposited with C3b, cf. intrinsic activity for DAF and probably MCP.

Recently, it has been demonstrated that CR2 can also act as a cofactor for factor I cleavage of iC3b to C3c and C3dg in vitro, albeit at a lower rate (Mitomo et al., 1987). CR2 is an integral membrane protein with mol. wt. 140 000. It is the C3dg receptor and the Epstein Barr virus receptor. The structural gene of CR2 has been cloned by cross-hybridisation using CR1 probe. It is highly homologous to CR1 and also maps to chromosome 1 at the q32 region (Weis et al., 1986; 1987).

Similar to that of the classical pathway, the alternative pathway C3 convertase can be further stabilised by substituting Ni^{2+} for Mg^{2+} during the assembly of the convertase. It can also be stabilised by the alternative pathway nephritic factor, which is an autoantibody (Fishelson and Muller-Eberhard, 1982). Thus it is apparent that the convertases of the two pathways of activation are analogous in many aspects, such as the analogous functions of C4 and C3, C2 and factor B and Mg^{2+} in the assembly/activation process, and in the homeostatic control of the convertase activity by the analogous/identical controlling proteins such as C4bp, factor H, CR1, DAF and MCP. However, there is a number of characteristic features in the alternative pathway. It is initiated continuously at a low level by the 'tick-over' mechanism (Lachmann, 1979; Lachmann and Hugh-Jones, 1984) as C3b-like (or C3(H2O)) molecules are generated spontaneously by slow hydrolysis of native C3 and these C3(H2O) molecules can assemble with factor B to form a C3 convertase after cleavage of factor D (Pangburn and Muller-Eberhard, 1980; Fishelson et al., 1984). Therefore, activators of the alternative pathway can be viewed as being able to accelerate

the formation of C3b or slow down its breakdown. This is in contrast with that of the classical pathway as it is only initiated after the activation of the C1 complex by antibody-antigen complex or other activators. Moreover, factor D which cleaves C3bB to form the alternative pathway convertase, C3bBb, circulates in plasma in an active form, unlike the activation of C1s by C1r after an autoproteolytic activation process in the C1 complex. A second major difference is the presence of a positive feedback mechanism in the alternative pathway as the C3b produced by C3 convertase is also a subcomponent of the same convertase. C3b can assemble with factor B to form more convertase. This results in the formation of a very efficient amplification loop.

1.2.4 The membrane attack complex (MAC)

Proteolysis of C5 to C5a and C5b by either C5 convertases is the last step of the activation process, but is the initial step of the lytic pathway. Altogether there are five components involved, i.e. C5b, C6, C7, C8 and C9. These components assemble sequentially into a macromolecular membrane attack complex, C5b-9, which will cause lysis of a variety of target cell types extending from viruses and bacteria to platelets and nucleated mammalian cells. These complement components are hydrophilic glycoproteins in the plasma, but during the spontaneous assembling process, they undergo a transition to an amphiphilic state that enables them to insert into the lipid bilayer of the target cell membrane. The result is the formation of 'doughnut-like' plugs which are hollow protein cylinders that transverse the target membrane

to generate hydrophilic transmembrane channels. The free permeation of ions, water and macromolecules through such channels, resulting in disruption of transmembrane gradients, may all contribute to the lethal effect of the MAC.

Similar to C3 and C4, C5 is synthesised as a single chain precursor. Before secretion, it is processed to two disulphide linked polypeptide chains α and β with 115,000 and 80,000 mol. wt., respectively. Activation of C5 generates the anaphylatoxin C5a, in addition to the major fragment C5b. C5a is the most potent complement anaphylatoxin and a leukocyte chemotactic peptide (reviewed in Hugli, 1986). The biological activities of C3a, C4a and C5a are controlled by anaphylatoxic inactivator, which acts like a carboxypeptidase B in removing the C-terminal arginine from the activation peptides. C5b in its nascent state constitutes the nucleus in the assembly of MAC. It possesses a metastable binding site with specificity for C6, which is probably due to transient expression of hydrophobic sites. Electron microscopy of negatively stained C5 revealed a multilobal, irregular molecular structure with estimated dimensions of 168 X 151 X 104 Å (Discipio et al., 1983). The primary structure of human C5 has been partially elucidated (Lundwall et al., 1985). Derived amino acid sequence covering the C-terminal of the chain and the N-terminal of the chain confirms that C5 is markedly homologous to C3, C4 and α 2 macroglobulin, although C5 lacks a thiolester bond compared with the other three related proteins. The structural gene of mouse C5 has been assigned to chromosome 2 (D'Eustachio et al., 1986).

C6 and C7 are single chain glycoprotein molecules with mol. wt. ~120,000 and 110,000, respectively. Family studies of genetic polymorphisms of C6 and C7 suggested that these two genes are closely linked but their chromosomal locations are still unknown. They also exhibit similar physical and chemical properties, and it has been suggested that they may be evolved from a common ancestral gene (Lachmann and Hobart, 1978; 1985). The nascent C5b in association with C3b (and probably C4b) firstly binds to C6 forming a C5b-6 complex and then to C7 to form C5b-7 complex. The C5b-7 complex appears to undergo a hydrophilic-amphiphilic transition upon C7 binding, thus generating a metastable binding site for membrane Anchorage of the C5b-7 complex is probably through high-affinity phospholipid binding site that may be located in the C7 molecule. Membrane-bound C5b-7 does not harm target cell but makes the target cell ready for further attack. It constitutes the receptor for C8 β that binds and processes C8 so that its α chain can enter the target membrane. C5b-8 produces a small membrane pore sufficient to lyse erythrocytes but less efficient to kill nucleated cells.

Human C8 is a structurally complex component that consists of equimolar amounts of three non-identical subunits, α , β and γ with 64,000, 64,000 and 22,000 mol. wt., respectively. These are arranged as a disulphide-linked α - γ dimer that is non-covalently associated with β . The α subunit contains the binding site for C9 and some undefined structure to interact with the target membranes. The β subunit has a domain that specifically mediates recognition and binding

of C8 to C5b-7, and also a region that interacts with target membranes. No function has been ascribed to the γ chain of C8.

The membrane-bound C5b-8 complex is a receptor for C9 and a catalyst for the polymerisation of C9 in the formation of membrane lesions. The C5b-9 complexes are heterogeneous with respect to their C9 content that may vary between 1-18 molecules per C5b-8 complex, depending on the availability of C9 monomer. Polymerisation of C9 can also be induced in vitro from high concentration of purified C9 on its own at 37°C, neutral pH in the presence of 50 mM Zn^{2+} (Tschopp et al., 1982). This forms tubular structures containing between 12-18 C9 molecules that resembles morphologically to the MAC that was firstly described by Borsos and coworkers (1964). It is now generally accepted that circular lesions produced by complement lysis are polymerised C9, although it is known that the C5b-9 does not necessarily manifest itself as the typical ultra-structural membrane lesion.

C9 consists of a single polypeptide chain of mol. wt. ~70,000. The monomeric C9 can be cleaved by α -thrombin to form the N-terminal C9a (mol. wt. 34,000), and the C-terminal C9b (mol. wt. 37,000) that is hydrophobic and is the portion of C9 that inserts into the phospholipid membranes. It is proposed that when C9 binds to C5b-8, it undergoes a major conformational change in which the molecule approximately double its length and the hydrophobic structure becomes exposed. The latter is also associated with the expression of neoantigens (Podack et al., 1982).

An unexpected but reasonable evolutionary relationship

between C8 and C9 was recently unravelled after cDNA clones coding for C9 (DiScipio et al., 1984; Stanley et al., 1985), C8 α (Rao et al., 1987), and C8 β (Haefliger et al., 1987; Howard et al., 1987) were obtained. An overall homology in primary structure of 24-33% identity has been found among these three proteins. These proteins have similar number of amino acid residues (C8 α , 552 residues, C8 β , 536 residues and C9, 537 residues). In addition, all of these molecules at the N-terminal region contain the class A, negatively charged Cys-rich 40-residues repeat which is a characteristic of the LDL receptor, and at the C-terminal region, the class B, Cys-rich 30-50 residues repeat which is a characteristic of the epidermal growth factor. Thus it has been postulated that C8 α , C8 β and C9 are evolved from the same ancestral gene. As mentioned, all these proteins are originally hydrophilic, but upon the binding of C5b-7 or C5b-8, they exposed the membrane-interacting segments by conformational changes. Therefore, C8 α , C8 β and C9 are members of a family of proteins that are capable of induced conformational changes leading to membrane interaction. To this end, a fourth member of this gene family could be preforin, which is a C9-related protein found in cytotoxic T cells and may be involved in cytotoxic killing (Lachmann, 1983; Young et al., 1986).

Genetic data suggests that human C8 α and C8 β are located on chromosome 1 (Pericak-Vance et al., 1982; Rogde et al., 1985). It was initially considered that C8 α and C8 γ are encoded by the same locus but it has been found that they are actually encoded by separate locus. C8 γ cDNA has been cloned and its mRNA is ~1 kb in size (Ng et al., 1987). The

chromosomal location of C9 is not known.

S protein is the single known controlling protein of the MAC. Binding of S protein to fluid phase C5b-7 may prevent its attachment to cell membranes. Although C8 and C9 bind to SC5b-7, the cytolytic terminal complex containing polymerised C9 is not formed. Therefore, the function of the S protein may be the solubilisation of fluid phase C5b-9 complexes, the protection of the bilayer cell by fluid phase C5b-7 and inhibition of C9 polymerisation during fluid phase assembly. cDNA clones of S protein show that it consists of a single polypeptide chain of 459 amino acid residues (Jenne and Stanley, 1985).

In conclusion, many complement proteins shared various degrees of sequence homologies which infer that they may possess similar structural or functional protein modules / domains. Most components of the complement system seem to have evolved from many gene duplication events although relatives for some proteins, e.g. properdin and C8 γ have not yet been found. Furthermore, some complement components, such as factor B and factor H, have internal sequence homologies. These are likely to have been formed by exon shuffling and/or by exon duplication events (see Table 1.2 and section 1.3 for further description).

1.3 Correlation of exon structure and protein structure / function

The application of DNA sequencing techniques to the study

Table 1.2 Structural kinships among complement proteins

A). Gene duplication*

1. C1q A, B, C chains
2. C1r, C1s
3. C2, factor B
4. C3, C4, C5 (α_2 macroglobulin, pregnancy zone protein)
5. C6, C7
6. C8 α , C8 β , C9 (perforin)

B). Exon shuffling/duplication+

- | | | |
|----|------------------------|--|
| 1. | SCR | factor B (3), C2 (3), C1r (2),
C1s (2), DAF (4), CR1 (~30),
CR2 (~15), C4bp (8),
factor H (20), MCP (?) |
| 2. | Serine protease domain | factor D, factor B, C2, C1r,
C1s, factor I |
| 3. | EGF Cys-rich module | C8 α , C8 β , C9, C1r, C1s, factor I |
| 4. | LDL Cys-rich module | C8 α , C8 β , C9 |
| 5. | Collagen-like domain | C1q A, B, C |

C). Others

1. Serine protease inhibitor family: C1 inhibitor
2. Carboxypeptidase (N): anaphylatoxin inactivator
3. Substrate adhesion protein family: S protein
4. Unknown origin: properdin, C8 γ

* Structural homologies extend throughout the entire protein molecules. Proteins which are not members of the complement system are enclosed in brackets.

+ Structural homologies are located in part of the protein molecules only. The number of repeating units found is enclosed in brackets.

of eukaryotic genes quickly established that most genes of eukaryotes are mosaic of expressed sequences and unexpressed, silent DNA. The former is defined as the exon, and the latter is the intron (Gilbert, 1978). Comparison of genomic DNA sequence of chicken ovalbumin (Breathnach et al., 1978) and chicken ovomucoid (Stein et al., 1980) genes with their respective corresponding mRNA sequences established the consensus sequence of 5' and 3' splicing sites of introns to be GT-AG respectively. Introns can be classified to three classes according to their position in the triplet codon at the exon/intron boundaries. The class 0 introns interrupts the reading phase between codons, class 1 introns interrupt codons between the first and second nucleotides and class 2 introns interrupt codons between the second and third nucleotides (Sharp, 1981). Except for the first and the last exons, there are two exon/intron splice junctions for each exon. Therefore, exons can be classified to 9 phases according to their corresponding intron phase at each end. During the past ten years the structure of numerous eukaryotic genes have been elucidated. The size of introns appears variable which ranges from 50 to 10,000 to 20,000 bp in length. However, the distribution of exon size is rather narrow and peaks at 40 to 50 amino acid residues (Blake, 1983; Naora and Deacon, 1982). Comparison of the intron positions among homologous genes in different species suggests that there is a temporal stability of exon/intron pattern. For example, in human, rabbit and mouse globin alpha chains, the introns are all located within the codon for residue 31 and between codons for residues 99 and 100. A similar pattern has also been observed in the beta

chains of globin (reviewed in Blake, 1985). On the other hand, there is indirect evidence revealing that introns can be lost and that exon/intron boundaries sometime migrate during the course of evolution. Typical examples can be found in the collagen genes and the genes encoding serine proteases.

The α -2 pro-collagen gene in chicken is composed of at least 52 exons, majority of them are 52 bp long, others are exactly twice as long (108 bp) and a few are 9 bp shorter than these (i.e. 45 and 99 bp). This is thought to be due to the precise deletion of the intervening sequences between adjacent 54 bp units that would lead to the double-length exons, and the deletion of 9 bp from the 54 and 108 bp exons (deCrombrughe and Pastan, 1982).

Comparison of exon/intron splice junctions among many serine proteases and the dihydrofolate reductase genes from various species suggests that the splice junctions correspond frequently to regions of polypeptide chains that show length variations. Thus it was suggested that extension or contraction of exons at the intron junctions constitute a mechanism for generating length variation and divergent sequences in protein families (Craik et al., 1983). Furthermore, correlation of splice junctions with protein structures suggests that splice junctions frequently map to protein surfaces (Craik et al., 1982; 1983).

Many of the vertebrate plasma proteins are homologous in structure and they are likely to be due to duplication of the same ancestral genes. For example, the complement component

genes C3, C4 and C5, the protease inhibitor α_2 macroglobulin and the pregnancy zone protein share some common biochemical properties and protein sequence homology. It has been suggested that these proteins originated from the same ancestral gene by gene duplication, followed by divergent evolution. Novel proteins can also be generated by exon shuffling events, which includes duplication and reassortment of exons from different sources. This is most obvious in proteins with repeating domains. Thus the 52 exons of the α_2 pro-collagen gene is probably due to the many duplication events of a single 54 bp ancestral exons. The ~60 amino acid repeating unit characterised in C4b/C3b binding proteins, SCR, mentioned in the previous section may also be derived from the same ancestral exon. A similar phenomenon probably also occurred in members of the Ig superfamily which might all derived from an ancestral exon coding for the characteristic disulphide-linked structure known as the immunoglobulin fold.

Assortment of blocks of exons from different genes to generate mosaic, multi-domain proteins might also have occurred. For example, complement factor B contains a serine protease domain which is encoded by the eight exons corresponding to the carboxy-terminus of the Bb fragment; three SCR at the amino-terminal region of Ba encoded by exons 2 - 4; and the factor B specific region covering the C terminus of Ba and the N terminus of Bb. Thus this complement protein might have derived from at least three diverse sources (Campbell and Porter, 1983; Campbell et al., 1984). Further classical examples can be found in the genes coding low density lipoprotein (LDL) receptor, complement

component C9 (and C8). These proteins appear to have two types of cysteine-rich protein modules encoded by separate exons characteristic of EGF and LDL receptor, respectively; and regions of protein derived from unknown origin (reviewed in Doolittle, 1985; Gilbert, 1985).

A striking feature in most of the repeating structures in vertebrate plasma proteins is that almost all of their corresponding exons are flanked by phase 1 introns. In other words, the exons are symmetrical in the sense that they have 1-1 boundaries at both ends (reviewed in Patthy, 1987). Obviously, symmetrical exons facilitate recombination events between intronic sequences. This is because tandem duplication, insertion or deletion of non-symmetrical exons would disrupt the reading frame and therefore may be deleterious.

As the split gene organisation of eukaryotes were discovered, it was soon hypothesised that exons encoded useful protein functions and that 'gene-in-pieces' nature would speed up evolution of eukaryotic genes through exonic modification, duplication and recombination (Gilbert, 1978). Thereafter many attempts have been made to correlate exon with protein structures and/or function. In view of the striking, direct correlation between the exon structure and the domain structure in the constant regions of immunoglobulin heavy chains (Sakano et al., 1979) and that the heme-binding domain of globins are encoded by a single exon, Blake (1970, 1979) and Go (1981, 1983) proposed that genes-in-pieces might suggest proteins-in-pieces. In other words, exons might

correspond to protein domains, or delineate compact units of protein structure. As more structures of eukaryotic genes were elucidated, it has become evident that this may not necessarily be the general case, as many single domain polypeptide chains are encoded on a number of exons (Philips et al., 1983). On the other hand, many of the repeating units being exchanged, such as those in plasma proteins, are relatively small and usually contain disulphide bonds, and are generally encoded by a single exon. Thus it does infer that some exons may delineate polypeptide fragments with certain folding. It has been suggested that the minimum size of peptides that can assume a stable folding structure is in the range of 20 - 40 residues (Wetlaufer, 1981). Incidentally, this range coincides with the size of exon-encoded protein fragments which peaks around 40-45 residues. As a result, Blake (1983, 1985) revised their notion and proposed that exons may correspond to sequential supersecondary structure in proteins. These oligopeptides probably form the basic elements on which larger and multiple domain proteins were assembled during the course of protein evolution. Although it is still difficult to generalise the correlation between exon and protein structure or function, it is quite clear now that exons code for segments of proteins that can be sorted independently during evolution. These segments of proteins can be related to functional regions, folding elements, domains or sub-domains (Gilbert, 1985).

1.4 Complement component C4

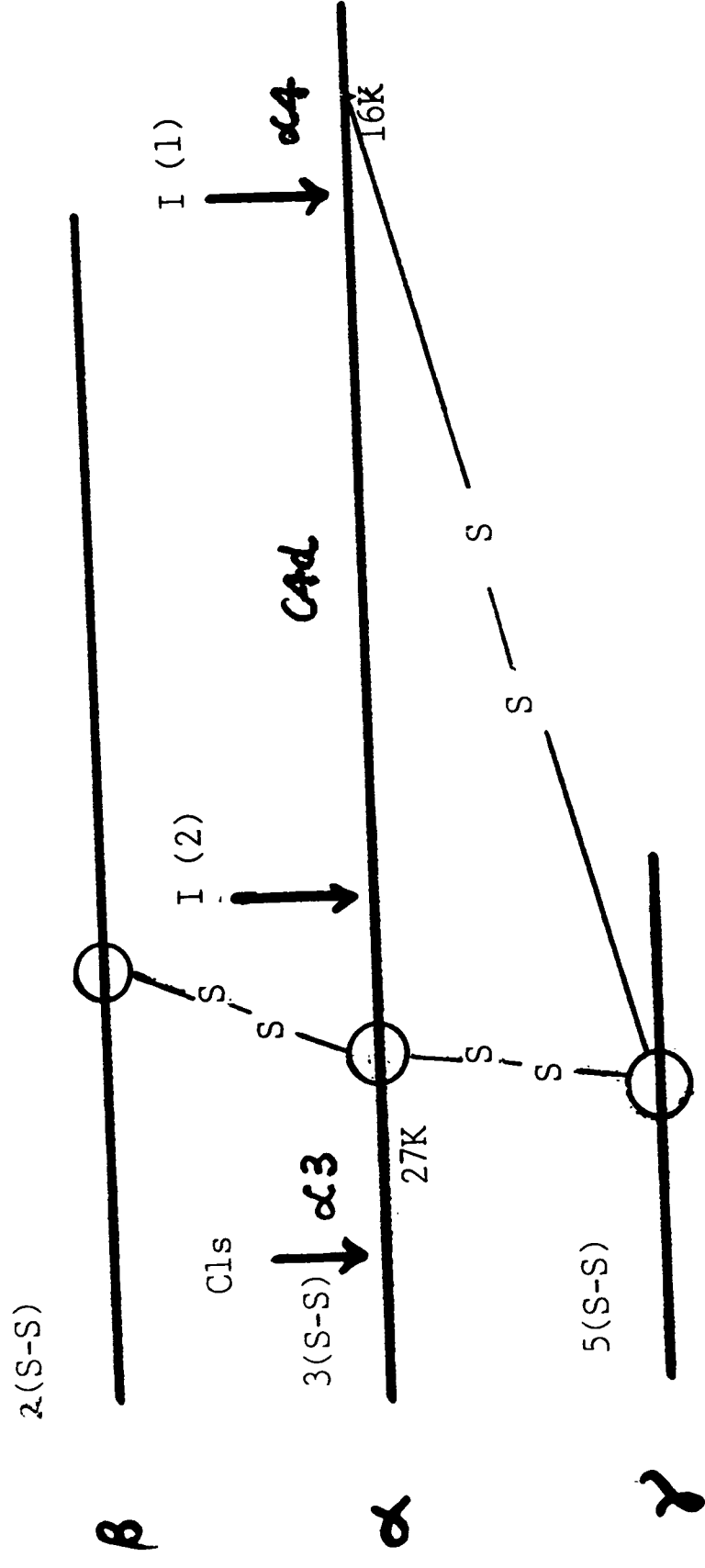
1.4.1 Biochemistry of C4

As mentioned in section 1.2, C4 is essential for the formation of the classical pathway C3 and C5 convertases. It is activated by C1s, and inactivated by factor I with C4bp or CR1 or MCP as cofactors, both through proteolytic cleavage(s) on the α chain of the disulphide linked three-chain molecule. Activation of C4 generates C4a, which resembles C3a and C5a, are anaphylatoxins. It also produces the metastable C4b, that has the ability to bind covalently to nearby cell-surfaces. This is due to the transacylation reaction from a thiolester bond that becomes labile upon activation of the C4 molecule. The nascent C4b has binding sites for C2 (C2a and C2b), C4bp, CR1 (and MCP, DAF?), C3 and probably C5. Some of these (non-covalent) binding sites may overlap. Their nature and location are still unknown. Inactivation of C4b by factor I is through a single cleavage to generate iC4b if MCP is used as a cofactor, or two cleavages to generate C4c and C4d if C4bp or CR1 is a cofactor (Fig. 1.5). On the other hand, inactivation of the nascent C4b can also occur by nucleophilic attack by water molecules that probably hydrolyse the intrachain thiolester linkage.

The major sites of biosynthesis of C4 include the liver and macrophages. The former probably constitutes the major source of plasma C4. Precursor C4 (pro-C4) is a single chain molecule, but is processed to the disulphide-linked three chain molecule before secretion (Schreiber and Muller-Eberhard, 1974; Gigli et al., 1977; Hall and Colten, 1977;

Fig. 1.5 Disulphide linkages, C1s and factor I cleavage sites
on human C4

There are one β - α and two α - γ interchain disulphide bonds on C4. On the α chain, Cys 801 and Cys 857 of the α 3 fragment are involved in the formation of a β - α and a α - γ disulphide bond. The Cys 1375 of the α 4 fragment is involved in a α - γ disulphide linkage. On the activation peptide C4a, there are 3 intrachain bonds which probably form a disulphide knot. There are two intrachain disulphide linkages on the β chain, and five on the γ chain. Their exact locations have not been resolved. O-S-S-O, interchain disulphide, positions not resolved; I(1), first factor I cleavage site; I(2), second factor I cleavage site; C1s, C1s cleavage site. (Modified from Janatova, 1986; and Seya et al., 1986).



Goldberger and Colten, 1980). However, this process is incomplete and the plasma contains some single chained pro-C4, and two chained $\beta + \alpha - \gamma$ and $\beta - \alpha + \gamma$ molecules (Gigli, 1978; Chan and Atkinson, 1983). In the plasma, native C4 undergoes a post-secretory cleavage such that the mol. wt. of the α chain is reduced by ~5,000 (Chan et al., 1983). The site of this cleavage is located 22 amino acid residues from the carboxyl terminus of the α chain (Law and Gagnon, 1985; Hortin et al., 1986a). Again, this process is incomplete and ~8% of the plasma C4 remains in the secretory form (Chan et al., 1983). Thus the C4 protein from a single gene locus can assume four different forms of expression (see Fig. 1.6). Normally, an individual possess three to five C4 genes which may be different due to polymorphism. Therefore, it is possible that there exists about 20 'faces' of C4 in some individuals (Chan et al., 1984). All these completely or incompletely processed forms of C4 are indistinguishable in terms of haemolytic activity.

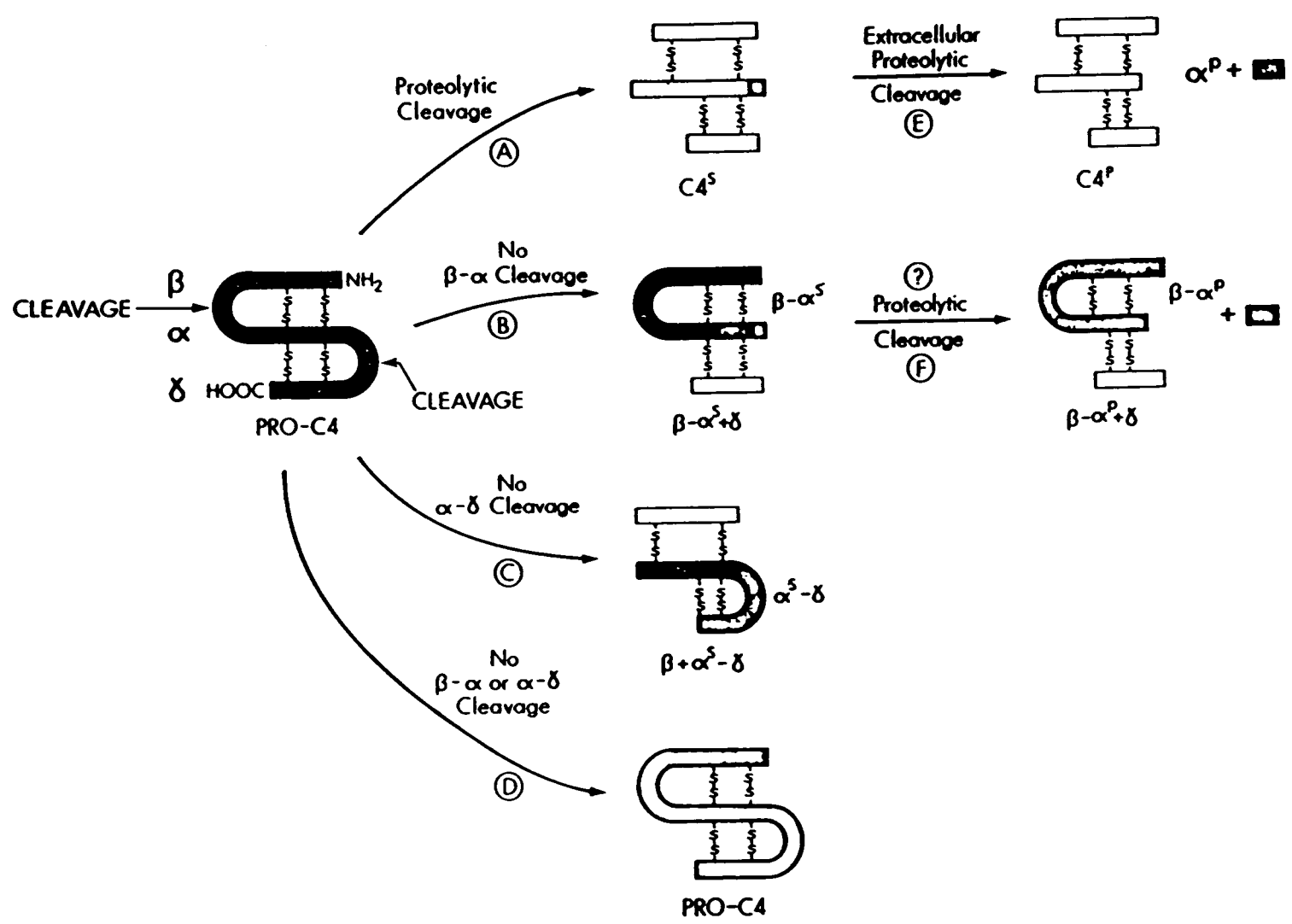
Besides proteolytic modifications, the C4 protein molecule also undergoes many other post-translational processing events. These include sulphation on tyrosine residues, addition and maturation of multiple oligosaccharide chains and formation of an internal thiolester bond.

The presence of tryosine-O-sulphation is a unique property of C4 among the complement proteins. This was firstly identified by Karp (1983). The site(s) of sulphation was located to the carboxyl end of the α chain. Further studies on the pattern of sulphation of C4 produced by HepG2 cells

Fig. 1.6 A model of the processing of the C4 precursor

- A. Cleavage at the β - α and δ - γ junctions to yield the predominant, 'secreted' three-subunit molecule.
- B and C. Cleavage at only one site to yield secreted two-subunit molecules with either uncleaved β - α or α - γ chains.
- D. Secretion of the uncleaved precursor.
- E. Extracellular cleavage of a C-terminal fragment to yield the predominant plasma form of C4.
- F. The probable extracellular processing of the uncleaved β - α chain.

(After Chan et al., 1984).



identify three sulphated tyrosine residues at position 738, 741 and 743 on the α chain (see chapter 3 for numbering from pro-C4). The region of sulphation on C4 is surrounded by many negatively charged amino acid residues which is also a property shared by other tyrosine-sulphated proteins such as fibropeptides and gastrins (Hortin et al., 1986b). No biological function for the sulphation of C4 has yet been discovered, although it was suggested that they might be related to the binding of C4bp (Karp, 1983) or to Mg^{2+} (Hortin et al., 1986b).

Approximately 7% of the total weight of human C4 is due to carbohydrate (Gigli et al., 1977). The oligosaccharide structure of C4 from plasma and HepG2 cells has been studied in considerable detail by Chan and Atkinson (1985). It was found that both of the β and α chains are glycosylated, whereas the γ chain is devoid of carbohydrate. The α chain has three complex fucosylated oligosaccharides of the biantennary type, with one each on the $\alpha 2$, $\alpha 3$ and $\alpha 4$ fragments. The β chain has a single high-mannose-oligosaccharide-type of glycosylation. The partially and completely processed C4 due to proteolysis have the same glycosylation pattern. Human C4 protein secreted by macrophages which has been incubated with tunicamycin to inhibit glycosylation did not alter their haemolytic activity, as compared with normal C4 molecules (Matthew et al., 1982).

On activation of C4, a labile group is exposed in the α chain of C4b that enables the molecule to bind covalently to a variety of surfaces by formation of an ester or an amide bond

(Law et al., 1980a; Campbell et al., 1980). At the same time, a free sulphdryl group is found that is undetected in the native molecule (Janatova and Tack, 1981). On the other hand, small nucleophiles such as methylamine can covalently bind to native C4 with concomitant appearance of a free sulphdryl group. The identical phenomenon is also observed in complement C3 and λ 2 macroglobulin. Double labelling experiments using radioactive methylamine and radioactive iodoacetamide (to label the free SH group) together with amino acid sequencing of C3 established that the free sulphdryl group is derived from a cysteine residue, while methylamine covalently bind to a glutarate residue which is only three residues away from the labile cysteine residue (Tack et al., 1981). Thus it was suggested that a thiolester bond is present between the β Cys- γ Glu residues. A similar conclusion was also reached from works on λ 2 macroglobulin (Sottrup-Jensen et al., 1980). Since C4 also shares the same amino acid sequence (Campbell et al., 1981; Harrison et al., 1981) around the proposed thiolester region and it also exhibits the same chemical reactivities described, it is generally accepted that C4 also contains a thiolester bond (Law et al., 1980b; reviewed in Tack, 1985; Reid and Porter, 1981; Porter, 1984).

There are some unique features for these thiolester containing proteins. One is the autolytic cleavage, which specifically splits a peptide bond on the λ chain to two fragments, by protein defaturants such as heat or SDS (Sim and Sim, 1981). The second one is the lability of the 'exposed' thiolester bond in activated C4b and C3b (half life at 37°C = 0.1 - 1 millisec). This is in marked contrast with simple

thiolester bond formed in synthetic peptides (half-life = 30 min at 37°C), and with native C4 and C3 (half-life = 3 months at 0°C) (Sim et al., 1981). Normally about 90% of the nascent C4b (and C3b) will be inactivated through hydrolysis by the surrounding water molecules (Law, 1983), while a minor proportion will covalently bind to the surfaces of the complement activator through a transacylation reaction to form amide or ester bonds. Recently, it has been shown that the two isotypes of C4, C4A and C4B, exhibit markedly differential binding affinity to amino groups and to hydroxyl groups of substrates (Isenman and Young, 1984; 1986; Law et al., 1984a; Dodds et al., 1986). On comparison, C4A has a strong preference to form amide bond while C4B has a higher affinity to form ester bonds. It is hypothesised that the differential thiolester activities is modulated by the isotypic residues of C4A and C4B (Yu et al., 1986; chapters 4 and 5).

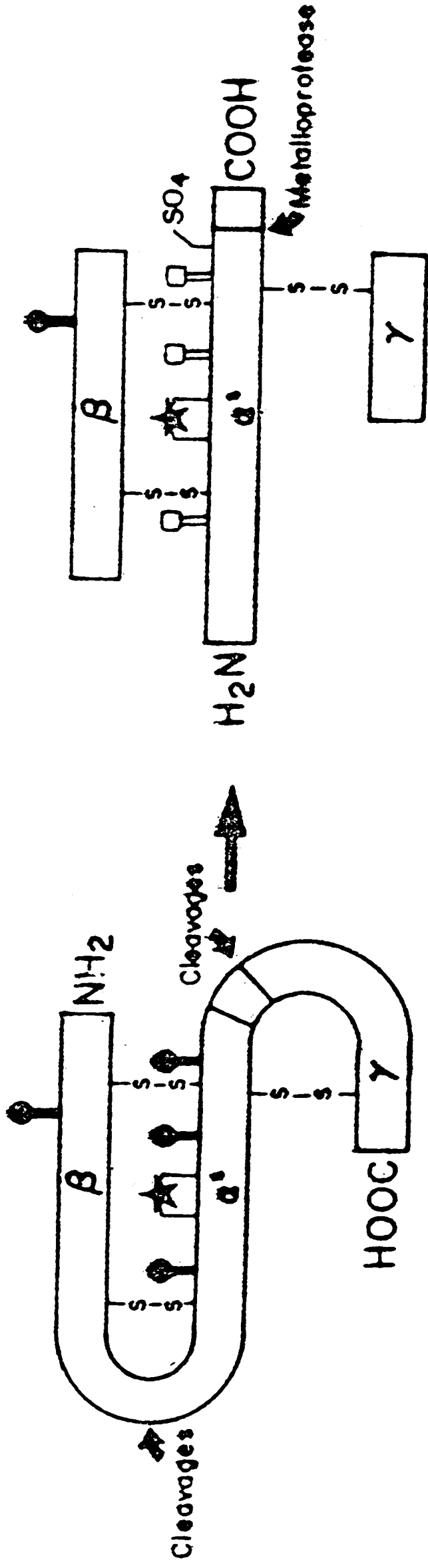
The activation peptide C4a, like C3a and C5a, promotes contraction of the smooth muscle, increases vascular permeability and release of histamine from mast cells and basophilic leukocytes. However, this anaphylatoxic activity from C4 is effective in the micromolar range, making it 100 - 1,000 fold less potent as a spasmogen than either C3a or C5a. Similar to C3a, C4a fails to elicit a significant chemotactic response from human neutrophils, while C5a in the nanomolar range promotes chemotaxis (reviewed in Hugli, 1986). C4a shows marked sequence homology with the other two anaphylatoxins, C3a (30% identity) and C5a (36% identity). The complete amino acid sequence of C4a was determined by Moon

and coworkers (1981), and is 77 amino acid residues in length. There are 6 cysteine residues and all of their positions are conserved compared with C3a and C5a. This also applies to the C-terminal residues, i.e. Arg 77 for C4a and C3a, or Arg 74 for C5a, which is essential for the anaphylatoxic activity. Inactivation by the anaphylatoxin inactivator is achieved through the proteolytic cleavage of the C-terminal Arg residue. Thus it was proposed that C3a, C4a and C5a are homologous molecules and have evolved from the same origin (Gorski et al., 1979). The crystal structure of human C3a has been obtained (Huber et al., 1980). This allowed the precise assignment of the three intra-chain disulphide bonds and also the modelling for the structure/function correlation of other anaphylatoxins (Greer, 1985).

Thus the role of C4 in the complement system can be summarised as follows. It acts as a non-catalytic link protein in the classical pathway convertases. It localises the response to the immediate vicinity of the site of activation and provides a base for the attachment of the lytic components. It also participates in immune adherence and opsonisation reactions by binding to specific receptors thus promoting phagocytosis (Lachmann, 1979). A schematic representation for the structure of the C4 molecule is shown in Fig. 1.7.

The presence of the intra-chain thiolester bonds, the similar biological properties of the anaphylatoxins, the multiple chain structure, the comparable mol. wt. and other biochemical properties shared by human C4, C3, C5 and α_2M

Pro-C4 is synthesised as a single polypeptide chain of 1745 residues (see chapter 3) and is processed by removing the leader peptide, the short basic tetrapeptides between the β , α and γ chains to give the three chain structure. After secretion, a further hydrolysis of a peptide (22 residues) from the C-terminal end of the α chain occurs. Three N-linked, biantennary complex type oligosaccharide structures are found on the α chain. The single glycosylation site on the β chain is of high mannose type oligosaccharide. A thiolester bond is located at the middle of the α chain and three clustered sulphation sites are found near the carboxyl terminus of the α chain. The figure is after Chan and Atkinson, (1985). Note, the interchain disulphide linkages are incorrect; see Fig. 1.6 for details.



PRO-C4

C4³

- = High mannose type oligosaccharide
- = Biantennary complex type oligosaccharide
- ☆ = Thiolester bond
- SO₄ = Sulfate

(Table 1.3) together suggests that these proteins are evolutionarily related. Molecular cloning of the structural genes encoding these proteins and the determination of their primary structures further substantiate the notion. Sequence comparison using the computer programme DIAGON has revealed stretches of sequence homologies between C4 and C3, and between C4 and α_2 M (Fig. 1.8a, b).

1.4.2 Genetics of C4

The genes encoding human complement C4, C2 and factor B were mapped to the major histocompatibility complex, HLA (reviewed in Porter, 1983, 1984, 1985; Campbell et al., 1986). Cosmid clones overlapping C2, factor B, C4A and C4B were obtained in 1984 (Carroll et al., 1984a). Subsequently, it was shown that the duplicated genes corresponding to the steroid P450 21-hydroxylase (21OHase) are present 3' to the complement C4A and C4B genes, respectively (White et al., 1984, 1985; Carroll et al., 1985c). The single gene encoding C2 is ~18 kb in size (Bentley et al., 1985). The gene coding for the functionally homologous protein factor B is only 6 kb long and is split into 18 exons (Campbell et al., 1984). It has been shown that the intergenic region between the human C2 and factor B genes is only 421 bp long (Wu et al., 1987). The C4 and 21OHase genes are ~30 kb 3' to the factor B gene (Fig. 1.9). Similar gene orders had been shown in the S region of the H2 complex in mouse (Chaplin et al., 1983). Recently, the orientation of the MHC class III genes with respect to the class I and class II genes has been resolved by pulse field gel electrophoresis. In mouse, the 21OHase B, C4B genes are ~350

Table 1.3 Comparison of the biochemical properties among human C4, C3, C5 and α_2 M

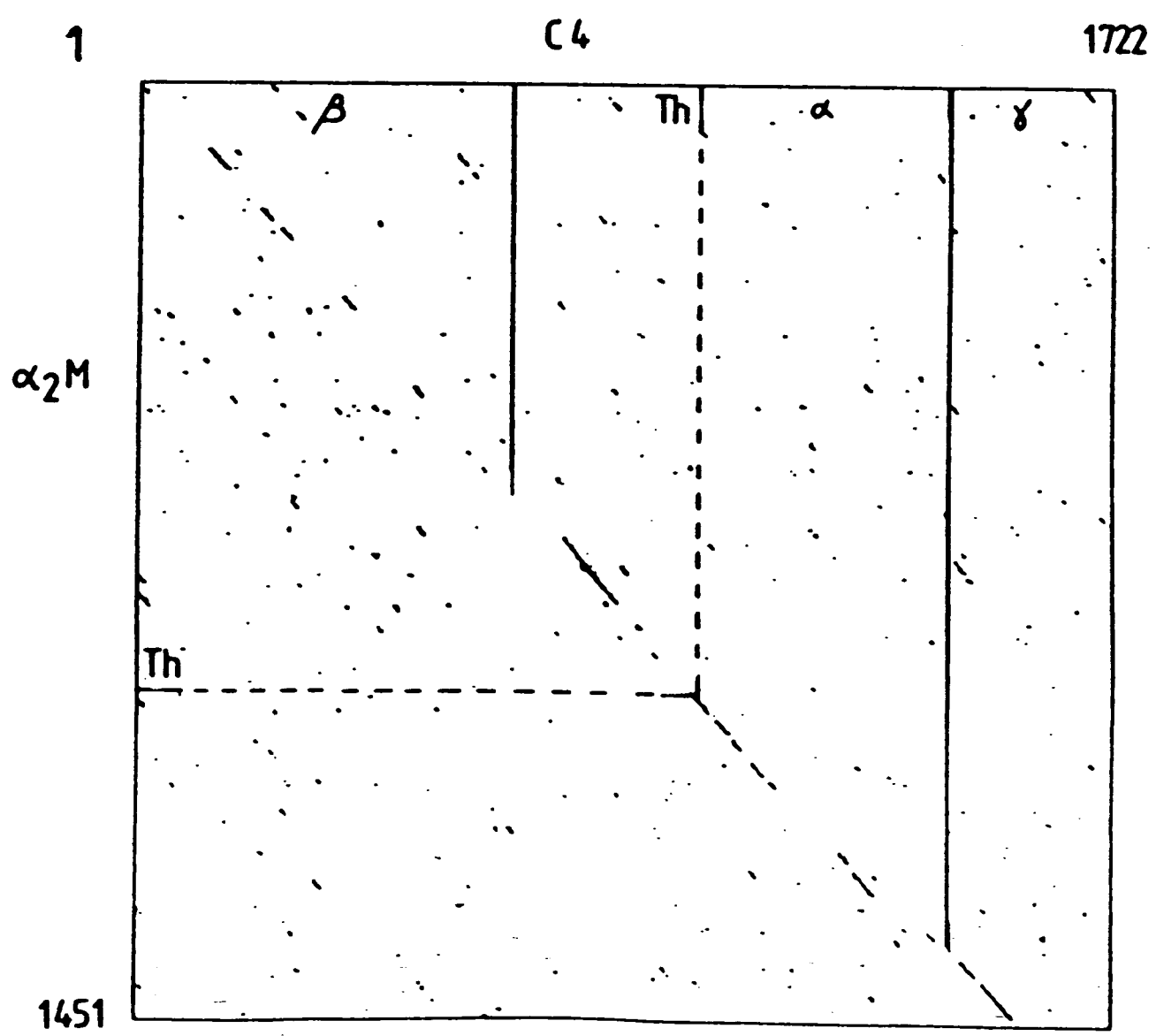
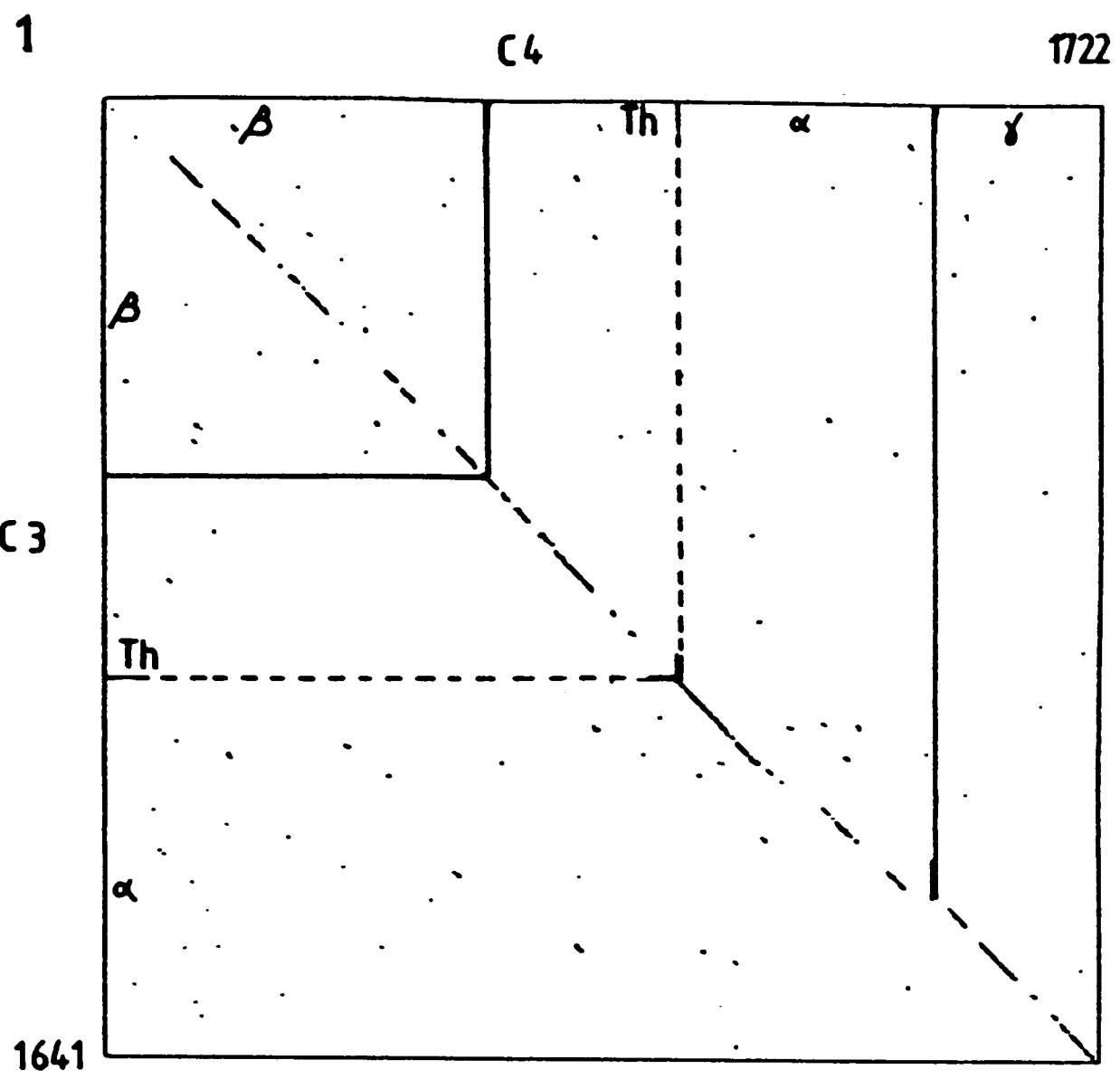
	<u>C4</u>	<u>C3</u>	<u>C5</u>	<u>α_2 M</u>
1. Chromosomal location	6	19	ND	12
2. Protein size (aa)	1726 ⁺	1641	ND	1451
mol. wt. (kd)	~200	185	195	180
3. Leader peptide (aa)	19	22	ND	23
4. Anaphylatoxin	C4a	C3a	C5a	Nil
size (aa)	77	77	74	-
5. No. of Cys	C4A: 28 C4B: 27	27	ND	23
6. Thiolester position	991-994* (1010-1013)	988-991*	Nil	949-952* (972-974)
7. No. of chains	3	2	2	1 (?)
(aa)	656	623	ND	-
(aa)	767	1015	ND	-
(aa)	291	-	-	-
8. RNA size (kb)	5.5	5.1	5.2	4.6

⁺ With size polymorphism

* The N-terminal amino acid residue is taken as number 1.
Brackets show the numbering as published.

Fig 1.8 Diagonal matrix comparisons between the amino acid
sequences of (a) pro-C4 and pro-C3, (b) pro-C4 and
 α_2 M

Homology among amino acid sequences of human C4, C3 and macroglobulin were analysed using the computer programme DIAGON (Staden, 1982b). The N-termini of both sequence pairs are at the top left corner and the total number of residues in each sequence is shown. The relative positions of the chains ($\beta - \alpha - \gamma$) and of the thiolester site (Th) are shown. The sliding window was set at twenty-one residues (After Terttia Belt, D. Phil thesis, 1985).



kb away from a class II gene E (Muller et al., 1987). A similar result has been obtained in the case of the human class III genes of the MHC and the orientation is established to be: centromere --- HLA D --- 21OHase B - C4B - 21OHase A - C4A - factor B - C2 --- HLA B --- HLA C --- HLA A --- telomere (Dunham et al., 1987; Carroll et al., 1987).

All of the HLA class III complement component genes and proteins are polymorphic (reviewed in Campbell and Bentley, 1985; Carroll et al., 1985a). Polymorphism of C4A and C4B at the protein level has been demonstrated by a variety of techniques based on differences in (1) electrical charges by agarose typing gels (Fig. 1.10; Mauff et al., 1983a; Sim and Cross , 1987); (2) apparent mol. wt. of the α chain (Roos et al., 1982) or the β chain (Mauff et al., 1983b) by SDS-PAGE; and (3) antigenic determinants by serological typing (Mauff et al., 1983a; Giles, 1987). Partial amino acid sequencing at the C4d region from pooled serum also detected variations in amino acid sequences (Chakravarti et al., 1983; Hellman et al., 1984). Nucleotide sequencing of partial and full length C4 cDNA clones for three different C4 allotypes has allowed the location of most of the polymorphic sites on C4, although it is difficult to correlate these structural variations with respect to the C4A/C4B isotypes and to the antigenic determinants. However, this laid down the foundation for further studies of C4 at the genomic level which has led to the elucidation of the variation of C4 in gene number (Carroll et al., 1985; Yu and Campbell, 1987), in gene size (Carroll et al., 1985a; Yu et al., 1986), the location of the C4A/C4B isotypic residues, and the Rg/Ch

Fig. 1.9 Alignment of the complement genes and the duplicated genes of 21-hydroxylase in the HLA class III region on human chromosome 6

The C2 and factor B genes are 421 bp apart (Wu et al., 1987). They lie approximately 30 kb from the C4A and C4B genes, which are separated from each other by about 10 kb. The class III genes lie between class I and class II genes. The C2 and factor B genes are closer to HLA-B, and the 21OH B, C4B genes are closer to HLA DR (Moller et al., 1987; Dunham et al., 1987; Carroll et al., 1987). The class III genes are overlapped by five cosmid clones as shown (Carroll et al., 1984a).

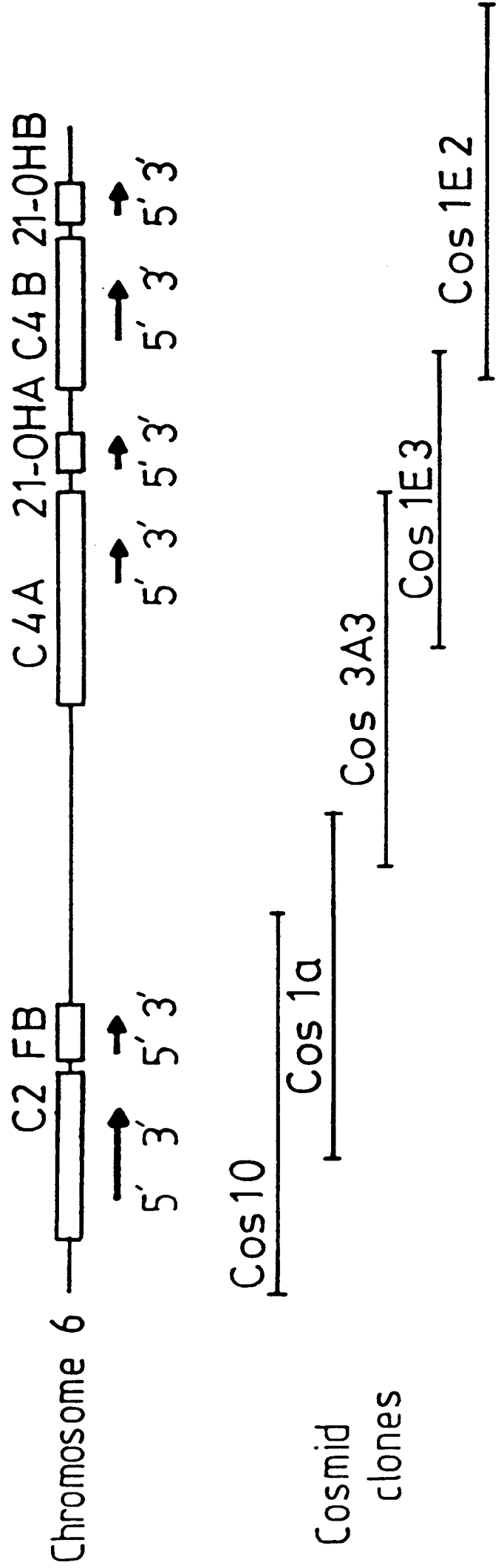
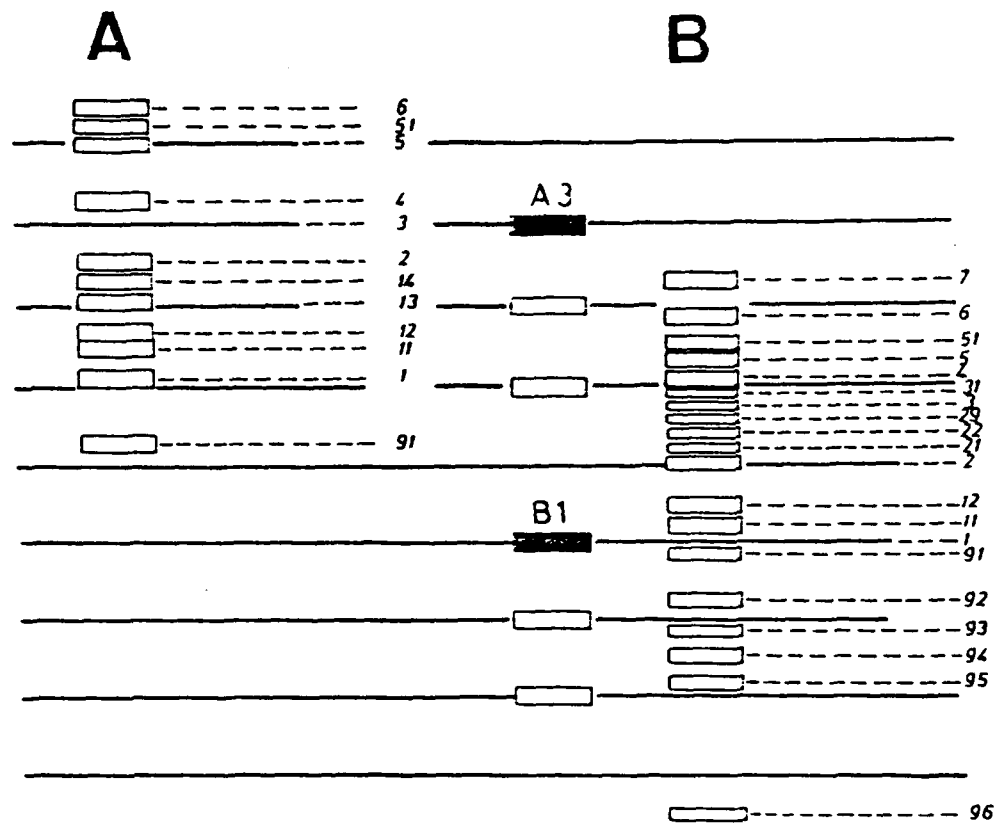


Fig. 1.10 Polymorphism of C4A and C4B based on the variation of electrophoretic mobilities

The electrophoretic pattern of various C4 allotypes are detected by agarose gel electrophoresis of desialised plasma, immunofixation with anti-C4 antibodies, and staining. Each allotype is represented by one major and two minor bands. This complicated band pattern has recently been simplified by Sim and Cross (1986), who used plasma pre-treated with carboxypeptidase B for typing. Presumably this treatment reduces the irregularities of the carboxyl termini of the β and the α chains. (Figure after Mauff et al., 1983).



Relative positions of electrophoretic human C4 allotype patterns (except for C4A 3 and B 1, only the major bands are shown).

antigenic determinants (chapters 4 and 5). The biological significance of C4 polymorphism is intriguing. It is of special interest to explore whether different polymorphic forms of C4 would vary in biological activities, and would correlate with HLA class I and II products. Another unusual phenomenon is the presence of some rare C4 alleles and null alleles with certain HLA class I and II alleles in biased combination (i.e. linkage disequilibrium) (Alper et al., 1983; 1986). More unexpectedly, individuals with some of these HLA haplotypes are more susceptible to autoimmune diseases (reviewed in Porter, 1983; Dawkins et al., 1983; Batchelor and McMichael, 1987). To resolve these questions, background knowledge on the genetics and functional activities of various forms of C4 in parallel with those of the HLA class I and II products will be essential.

CHAPTER 2

MATERIALS AND METHODS

2.1 Materials

2.1.1 Enzymes

Sources of enzymes are as follows.

a) Restriction Enzymes. All restriction enzymes were purchased from Amersham except the following.

Acc I, EcoO 109, Mnl I,	New England Biolabs
Nar I, Nhe I, Nla IV;	
Asp 718, Ban I, Ban II, Bgl II,	Boehringer Mannheim
Bst EII, Cla I, Dra I;	
Alu I, Nco I, Sst I, Xho I	BRL
Ban I, Dra I.	Pharmacia

b) Others

Lysozyme, RNase, DNase	Sigma
Alkaline phosphatase	Boehringer Mannheim
DNA polymerase I, Klenow fragment	Boehringer Mannheim, Amersham, Pharmacia
T4 polynucleotide kinase	Boehringer Mannheim
T4 DNA ligase	Gift from G.G. Brownlee Amersham
Nick-translation kit	Amersham
Multi-primed DNA labelling kit	Amersham

2.1.2 Chemical reagents

Chemical reagents were obtained from BDH, Sigma, Fisons and Boehringer Mannheim except for:

Agar No. 1, Yeast extract	-	Oxoid
Agarose	-	Seakem, Pharmacia and Sigma

Bactotryptone	-	Difco
Dextran sulphate, Ficoll	-	Pharmacia
Chemicals for oligonucleotide synthesis	-	Cruachem

2.1.3 Bacterial strains, plasmids, phages and cell lines

<u>Eschericia coli</u> strain	MC 1061	(Casadaban & Cohen, 1980)
"	"	"
"	JM 103	(Messing et al., 1981)
"	"	"
"	Q 358	(Karn et al., 1980)
"	"	"
"	Q 359	(" " " ")
Plasmids: pAT 153 Pvu II	- 8	(Anson et al., 1984)
pATX		
Phage: M13 mp8, mp9		(Messing & Vieira, 1982)
λEMBL-3		(Frischauf et al., 1983)

Media

2 x TY liquid medium : 15 g bactotryptone, 10 g yeast extract, 5 g NaCl to 1 litre with water.
pH adjusted to 7.4 with 2 ml 4M NaOH.

L broth liquid medium : 10 g bactotryptone, 5 g yeast extract, 10 g NaCl to 1 litre with water.

Solid medium : as above + 15 g/litre No.1 Agar.

T broth liquid medium : 10 g bactotryptone, 5 g NaCl to 1 litre with water.

Solid medium : as above + 11 g/litre No.1 Agar or Agarose (Pharmacia).

L-top and T-top Agar : Liquid media + 7 g/litre No.1 Agar
(Agarose) (Agarose)

2.1.4 Radioactive molecules

All radioactive molecules were obtained from The Radiochemical Centre, Amersham, Buckinghamshire, U.K.

[$\alpha^{32}\text{P}$]-dNTPs were 10 mCi/ml, specific activity 3,000 Ci/mmol.

[$\alpha^{32}\text{P}$]-dATPs were 10 mCi/ml, specific activity 5,000 Ci/mmol.

[$\alpha^{35}\text{S}$]-dATP was 8-10 mCi/ml, specific activity 1,200 Ci/mmol.

2.2 Overnight Cultures

Bacteria from a glycerol stock or from a single colony were used to inoculate 1-50 ml liquid media and incubated at 37°C overnight.

E. coli JM 103 were grown in 2 x TY broth

E. coli MC 1061 were grown in L broth.

E. coli Q 358 and Q 359 were grown in T broth/0.2% maltose and adjusted to 10 mM MgCl_2 just prior to use.

2.3 Preparation of DNA

2.3.1 Small scale isolation of plasmid DNA (10-100 ml) (Birnboim & Doly, 1979)

Cultures of E. coli were grown in L Broth containing 100 $\mu\text{g/ml}$ ampicillin to an A_{600} 0.8 OD units. 1 ml was removed and centrifuged at low speed in a MSE microfuge for 1 min at room temperature. The pelleted cells were resuspended in 200 μl Hogness freezing buffer (4% glycerol/3.6 mM di-potassium hydrogen phosphate/1.3 mM potassium dihydrogen phosphate/2 mM sodium citrate/ 1 mM MgSO_4 ; Hogness & Simmons, 1964) and stored at -20°C or -70°C. To the remaining culture chloramphenicol was added to 200 $\mu\text{g/ml}$ (Clewell & Helsinki, 1972) and the culture grown in a rotatory shaker at 37°C overnight.

The cells were pelleted in the MSE 6L centrifuge at 3000 r.p.m. for 10 min at 4°C, and then resuspended in 50 mM glucose/10 mM EDTA/25 mM Tris, pH 8.0 (150/10 ml culture). The cells were lysed by the following successive treatments (a) 5 mg/ml lysozyme for 30 min at 0°C, (b) two volumes 0.2 M NaOH/1% SDS for 5 min at 0°C, (c) one half volume 3 M NaAc, pH 4.8 for 30 min at 0°C. The solution was centrifuged twice at 11,000 g and the clear supernatant phenol/chloroform extracted. The plasmid DNA was precipitated from the aqueous phase with an equal volume of isopropanol, resuspended in 0.3 M NaAc (300 µl/10 ml starting culture) and reprecipitated with 2.5 volumes of ethanol. The pellet was washed with 70% EtOH, dried under vacuum and resuspended in 10 mM Tris/0.1 mM EDTA, pH 8.0 (20 µl/10 ml starting culture). The plasmid DNA was incubated at 37°C for 30 min with 5 µg/ml heat inactivated pancreatic RNase, and then stored at -20°C.

2.3.2 Large scale isolation of plasmid DNA, cosmid DNA (500-1,000 ml) (Radloff et al., 1967; Maniatis et al., 1982)

Cultures of E. coli were grown in L broth containing ampicillin but only plasmid DNA cultures were treated with chloramphenicol. After the overnight incubation the cells were harvested and lysed as described (Section 2.3.1). The phenol extraction step was omitted and the DNA was precipitated with isopropanol. The DNA pellet was washed, dried and processed either as (1) or (2).

1. Resuspend pellet in 15 ml of SET (150 mM NaCl/5 mM EDTA, pH 8.0, 50 mM Tris, pH 8.0), 300 µl of 0.5% ethidium bromide and an equal weight of CsCl were added to the DNA. The solution was transferred to a 38 ml Beckman ultracentrifuge tube, topped up with liquid paraffin, capped, sealed and centrifuged for 17 hr at 35,000 r.p.m. at 18°C in a Ti 60 rotor. The centrifuge was stopped without the

brake and the solution transferred away from the large RNA pellet to a clean centrifuge tube. This was capped as before and re-spun for 60 hr at 35,000 r.p.m. at 18°C.

2. Resuspend pellet in 6 ml of SET. 200 µl of 0.5% ethidium bromide and 8 g of CsCl were added to the solution. The solution was transferred to a 9 ml Beckman ultracentrifuge tube, topped up with liquid paraffin, capped, sealed and centrifuged overnight (~16 hr) at 55,000 r.p.m. at 18°C in a Ti 70 rotor (L8 centrifuge). The speed was then shifted to 45,000 r.p.m. and the rotor was spun for an hour under the same conditions so as to allow 'relaxation' of the DNA bands. The centrifuge was stopped.

Both procedure 1 and 2 achieved the similar result. Two bands of DNA were visible. The upper band of linear DNA and the lower band of closed circular DNA were collected separately by puncturing the tube with a needle just below the band.

Ethidium bromide was removed by repeated extractions with equal volumes of 1-butanol. The DNA was recovered by addition of 3 volumes water and 2 volumes ethanol. The DNA was reprecipitated once, washed, dried, resuspended in 10 mM Tris/0.1 EDTA, pH 8.0 and stored at -20°C.

The DNA concentration was estimated through spectroscopic measurement assuming that 1.0 OD₂₆₀ unit = 40 µg/ml of DNA. Normally, (intact) DNA extracted from the lower band was used for experiments.

2.3.3 Isolation of replicative form of bacteriophage M13 mp9

10 µl of an overnight culture of E. coli JM 103 was used to inoculate 1 litre of 2 x TY liquid broth and incubated at 37°C until A₆₀₀ = 0.6 OD units. 10 ml of an overnight culture of E. coli JM 103, infected with a single viral plaque of bacteriophage M13 mp9, were added to the culture and incubated at 37°C for a further 1½-2 hr (A₆₀₀ - 1.0). The

cells were then harvested and lysed as described (Section 2.3.1). The bacteriophage DNA was purified by centrifugation to equilibrium in CsCl gradients as described (Section 2.3.2). After ethanol precipitations the DNA was resuspended in 10 mM Tris/0.1 mM EDTA, pH 8.0 and stored at 4°C.

Replicative form of bacteriophage M13 mp8, prepared under the same method described above, was kindly given by K. T. Belt.

2.3.4 A small scale isolation of bacteriophage λ DNA (Manitatis et al., 1982)

About 10^5 p.f.u. of bacteriophage were mixed with 2 ml of E. coli Q 358 and incubated at 37°C for 30 min. This was mixed with 9 ml molten T top agarose, poured onto 14 cm T broth agarose plates and kept at 4°C for 4-8 hr. The phage suspension was transferred to a 50 ml Sorval SS34 centrifuge tube and kept at 4°C. The top agarose was scratched with a sterile spatula and transferred to a fresh SS34 centrifuge tube. 8-10 ml of phage buffer was added. The centrifuge tube was sealed and shaken at 37°C for 2 hr. It was spun at 8000 r.p.m. 4°C for 15 min, using a SS34 rotor in a Sorval RC-5B centrifuge. The supernatant was pooled to the original phage suspension and spun twice under the same conditions to remove residual bacterial debris. To the clear supernatant, pancreatic DNase I and RNase A was added, each at a final concentration of 10 μ g/ml and incubated at 37°C for 30 min. Polyethylene glycol (PEG) mol. wt. 8000 and solid NaCl were added to a final concentration of 10% (w/v) and 0.5 M respectively. The phage suspension in a fresh centrifuge tube was left at 4°C for 8-16 hr with occasional shaking. The bacteriophage particles were recovered by centrifugation at 10,000 r.p.m. for 20 min at 4°C in the Sorval centrifuge. The pellet was resuspended in 0.6 ml of phage buffer and left at 4°C for 2 hr.

Debris was removed by low speed spinning in a microfuge for 2 min. The supernatant, in two Eppendorf vials, was incubated in 0.1% SDS/5 mM EDTA pH 8.0 at 68°C for 15 min. The aqueous phase was extracted once with phenol, phenol/chloroform (1:1), 2-3 times with chloroform and then precipitated with an equal volume of isopropanol. The DNA pellet was washed, dried, resuspended in 50 µl 10 mM Tris/0.1 mM EDTA pH 8.0 and stored at -20°C.

2.3.5 Large scale preparation of bacteriophage DNA

6-10 plate lysates were prepared under the identical conditions described. 300 ml centrifuge bottles and MSE centrifuge was used. Speed of centrifugation was at 8000 r.p.m. After the PEG precipitation, the recovered phage particle was resuspended in 3 ml of phage buffer (in 6 vials). The supernatant was spun in microfuge to remove debris, and then chloroform extracted to removed PEG. It was further purified by spinning at 25000 r.p.m., 4°C, for two hr using a SW50 rotor in a Beckman L5 or L8 ultracentrifuge. The phage pellet was resuspended in 0.5 ml of phage buffer and kept at 4°C overnight. Residual debris was removed by low speed spinning in microfuge.

2.3.6 Isolation of chromosomal DNA (Bell et al., 1981)

Chromosomal (or genomic) DNA was isolated from EBV-transformed lymphoblastoid cells or from white cells (i.e. buffy coat) of blood. The buffy coat is a thin layer cell that forms between red cells and plasma when the whole blood is centrifuged at 2,500 g for 15-30 min.

Cell nuclei were prepared by adding 40 ml of lysis buffer (0.32 M sucrose/10 mM Tris, pH 7.4/5 mM MgCl₂/1% Triton-X100) to cells in a Falcon tube and spin at 2000 r.p.m. for 15 min at 5°C in the MSE 6L

centrifuge. The supernatant was discarded. This lysis procedure was repeated twice. The nuclei were resuspended in 10 ml SET (150 mM NaCl/5 mM EDTA/50 mM Tris, pH 8.0). To the suspension proteinase K was added to 100-200 µg/ml, SDS to 0.5%. It was incubated at 37°C until all of the clumps were solubilized. The solution was then extracted twice with phenyl: chloroform: isopropanol (25:24:1) and once with chloroform:isopropanol (24:1). After centrifugation at 2,000 r.p.m. for 15 min at 4°C the organic layer was removed using a syringe, leaving behind the aqueous layer and interface. For the final extraction the aqueous layer was removed to clean tube and genomic DNA precipitated with 2.5 volumes of chilled ethanol (-20°C). The DNA precipitated immediately and was removed with a sealed pasteur pipette. It was briefly washed in 70% ethanol, dried and resuspended overnight at 4°C in 200 µl 10 mM Tris/1 mM EDTA, pH 7.4. The DNA was then dialysed against 10 mM Tris/1 mM EDTA, pH 7.4 overnight at 4°C, and stored at 4°C at approximately 0.5-1 mg/ml.

2.4 Restriction Digests of DNA

Digestion of DNA by restriction enzymes were performed under manufacturer's suggested conditions. Analytic digests used 100-1000 ng DNA in a reaction volume of 10-20 µl. Preparation digests used 5-20 µg DNA in 50-100 µl reactions. Generally the amount of enzyme used was above the 1 unit/µg level. 1 unit of restriction enzyme is the amount stated by the manufacturer to fully digest 1 µg of DNA (mostly λ DNA) in 60 min. All digests were carried out at 37°C except Sma I (Boehringer Mannheim) at 25°C, Bst EII at 60°C and Taq I at 65°C. 30 µl of paraffin was added to the Bst EII or to Taq I digests. Genomic DNA was

restricted for 16-24 hr in a reaction volume of 40 μ l. All other DNA samples were fully digested after 2 hr.

2.5 Radiolabelling of DNA

2.5.1. 'Fill in' reaction (Wu and Taylor, 1971; Maxam and Gilbert, 1980)

α 32 PdNTPs were incorporated into DNA fragments by a 'fill-in' reaction using the 'Klenow' fragment of DNA polymerase I. Restriction digests were labelled directly by adding 5 units DNA polymerase I ('Klenow' fragment) per 50 μ l total reaction volume together with 1-50 μ Ci of α 32 PdNTP and dNTPs to a final concentration of 20 μ M (see below). The reaction was incubated at room temperature for 20 min. and chased with cold dNTP (corresponding to the α 32 PdNTP) at room temperature for a further 10 min.

For most 'fill-in' reactions the dNTPs added were the three necessary to complement the α 32 PdNTP. In the case of fragment preparation for blunt-end ligation into PAT-153 or M13, only 1/10 of digested DNA was end-filled with α 32 PdNTP as described. All four dNTPs (cold) were added to the rest to ensure that all the fragments were blunt-ended.

2.5.2 Kinase labelling (Maxam and Gilbert, 1980)

The 5' ends of oligonucleotides were labelled by mixing 25-50 ng DNA in 50 mM Tris, pH 7.6/10 mM MgCl_2 /10 mM DTT with 75 μ Ci [γ 32 P]-dATP and 8 U T4 polynucleotide kinase in a total volume of 20 μ l. The reaction was incubated at 37°C for 30 min then terminated with 20 mM EDTA. The DNA was desalted on a 1 ml Sephadex G-50 superfine column.

2.5.3 Nick translation (Rigby et al., 1977)

50-100 ng DNA in 5 mM MgCl_2 /50 mM Tris, pH 7.8/20 mM β -mercaptoethanol/50 $\mu\text{g/ml}$ BSA was mixed with 20 μCi [α ³²P]-dATP, 20 μM dCTP, dGTP, TTP and 5 μl Amersham 'nick translation' enzyme (2.5 U DNA Polymerase I and 50 pg DNase I) in a total volume of 100 μl . The reaction was incubated at 16°C for 2 hr then terminated with 10 mM EDTA and heating at 70°C for 10 min. The DNA was desalted on a 1 ml Sephadex S-300 column.

2.5.4 Multi-primed labelling (Feinberg and Vogelstein, 1984)

This technique involves the application of random sequence hexanucleotides to prime DNA synthesis on denatured DNA at numerous sites along its length. Restriction endonuclease DNA fragment can be radiolabelled to high specificity indifferent to inhibition by contaminants from agarose gels. 25-50 ng of DNA in 10 μl of sterilized water was denatured at boiling water bath for two min and then chilled on ice. 15 μl of commercial multiprime buffer solution (which contains dATP, dGTP, TTP random hexanucleotides, in a concentrated buffer solution with Tris pH 7.8, MgCl_2 , β -mercaptoethanol and BSA), 2 U of DNA polymerase I, Klenow fragment, 20 μCi α -³²PdCTP and sterilized water were added to a reaction volume of 50 μl . Radiolabelling was carried out at room temperature for 5 hr and terminated by adding EDTA to 20 mM. The DNA was desalted on a 1 ml Sephadex S-300 column. Normally more than 60% of label was incorporated.

2.6 Gel Electrophoresis

2.6.1. Polyacrylamide gel electrophoresis (Sanger and Coulson, 1978:

Maxam and Gilbert, 1980)

40 cm x 20 cm glass plates were used. The notched (i.e. top) glass plate was treated with a 2% solution of dimethyldichlorosilane in 1,1,1 trichloroethane, space with 0.3 mm strips of plastic card and taped

together with the bottom glass plate using waterproof electrical tape (i.e. Tuck Tape). Sample wells 2 mm-2.5 mm in width were formed at the top of the gel by insertion of a shaped plastic card 'comb' between the glass plates.

A 40% acrylamide stock solution was prepared by dissolving 38 g acrylamide and 2 g bis-acrylamide per 100 ml water final volume. The solution was de-ionised by stirring with 5 g 'Amerlite' MB-3 monobed resin for 30 min, filtered and stored at 4°C in a dark glass bottle.

4-20% gels were prepared by mixing stock acrylamide solution with 5 ml 10 x TBE (900 mM Tris/900 mM boric acid/25 mM EDTA, pH 8.3) in a total volume of 50 ml. 21 g urea were added for denaturing gels. To catalyse the polymerisation, 0.2 ml 10% ammonium persulphate and 0.08 ml TEMED were added, the solution mixed and poured. Loading dyes were added to each electrophoresis sample. For native gels a glycerol dye mix of 30% glycerol in 5 mM EDTA, 0.1% xylene cyanol and bromophenol blue was used. For urea gels a formamide dye mix of 60% formamide in 35 mM EDTA, 0.1% dyes was used, and all samples were boiled for 4 min then chilled on ice prior to loading.

Gels were electrophoresed in 1 x TBE for 1½-4 hr at a constant current of 15 mA for 20% gels and 20-30 mA for lower percentage gels. A thin metal plate was used to back up the bottom glass plate so as to circumvent the 'smiling effect' of the gel caused by uneven heating during electrophoresis.

2.6.2 Autoradiography (Laskey and Mills, 1977)

Gels were covered with cling-film and autoradiographed using X-OMAT 'S' X-ray film (Kodak) with a phosphotungstate intensifying screen (Cronex Lightening Plus, Dupont) at 4°C or -70°C. The films were sensitized by pre-flashing and developed in an MEI X-OMAT automatic processor.

2.6.3 Agarose gel electrophoresis (submerged gels) (Southern, 1979)

0.6-2% agarose horizontal slab gels were run using the BRL gel apparatus. 300 ml agarose in 1 x TBE were poured into the gel tray and left for 1 hr. 2 litres 1 x TBE were used as running buffer. Gels were stained by direct addition of 1 mg ethidium bromide to the agarose before pouring. With genomic DNA the gels were stained after electrophoresis by submerging the gel in 1 x TBE/2 mg/litre ethidium bromide for 30 min. Samples were mixed with glycerol dye mix and electrophoresed overnight at 50-70 mA.

For less accurate resolution, the ('Uniscience' minigel apparatus was used (Johnson and Grossman, 1977). 30 ml agarose was used to form the gel and running buffer was 60 ml 1 x TBE/10µl 0.1% ethidium bromide solution. Samples were mixed with glycerol dye mix and electrophoresed at 30-100 mA for 1-3 hr.

The ethidium bromide stained DNA (or RNA) was visualised under short-wave ultra-violet light (Sharp et al., 1973) using a 'Fotodyne' transilluminator, and photographed using a Polaroid MP4 camera fitted with a red filter.

2.7 Recovery of DNA from gels

2.7.1 Elution of DNA from acrylamide gels (Maxam and Gilbert, 1980)

The section of gel containing the DNA of interest (identified by superimposition of the autoradiograph on the gel) was excised with a razor blade placed in a siliconised glass tube containing 0.5 ml 2 M ammonium acetate. The DNA was eluted at 37°C overnight with constant shaking. After elution the supernatant was removed and the gel slice washed with 200 µl 2 M ammonium acetate. The supernatants were pooled and residual acrylamide removed by centrifugation for 5 min in a microfuge at room temperature. 1.5 volumes of ethanol were added to the supernatant and the DNA pelleted after precipitating at -70°C for 30 min. The pellet was resuspended in 300 µl 0.3 M NaAc, pH 6.0 and ethanol precipitated twice more before washing in 70% ethanol and drying

under vacuum. The DNA was resuspended in 10-30 μ l of 10 mM Tris/0.1 mM EDTA, pH 8.0.

2.7.2 Elution of DNA from agarose gels (Maniatis et al., 1982)

Dialysis membrane ('Visking') was boiled for 10 min in 2% NaHCO_3 /1 mM EDTA, rinsed, re-boiled for 10 min in distilled water, cooled and stored at 4°C in 20% methanol. To elute DNA from agarose minigels, an incision was made in the gel directly ahead of the fragment (visualised under ultraviolet light). A small piece of Whatman 3 mm paper backed by trimmed 'Visking' membrane was inserted into the incision and the gel electrophoresed for 2-3 min until the DNA fragment was no longer visible on the gel. The DNA was recovered by centrifuging the paper and the membrane in a 0.75 ml Eppendorf tube, pierced at the apex, nested in a 1.5 ml Eppendorf tube. The paper and membrane were washed twice with 100 μ l SET (150 mM NaCl/50 mM Tris/1 mM EDTA, pH 8.0) and the washing collected as above. The combined washing were extracted once with phenol/chloroform, ethanol precipitated, washed, dried and resuspended in 20-50 μ l 10 mM Tris/0.1 mM EDTA, pH 8.0.

2.8 Sub-cloning of DNA fragments into pAT 153/Pvu II/8 vector

2.8.1 Preparation of E. coli MC1061 competent cells

1 l of MC1061 cells were grown to an A_{600} 0.2-0.3 OD units, chilled on ice for 10 min and centrifuged at 3000 g for 10 min at 4°C in the MSE centrifuge using 6 x 500 ml rotor. The cells were gently resuspended in 0.2 volume 25 mM CaCl_2 /10 mM Pipes, pH 6.6 and transferred to two 50 ml Falcon tubes. The cells were re-pelleted in MSE-6L centrifuge at 2000 r.p.m. for 5 min and gently resuspended in 20 ml of 10 mM Pipes, pH 6.6/50 mM CaCl_2 /15% glycerol. 200 μ l aliquots were snap frozen in a dry ice/ethanol bath and stored at -70°C. The cells were thawed slowly on

ice for 10 min before use. Efficiency of transformation by pAT 153/Pvu II/8 vector was about $0.6-2 \times 10^7$ transformant per μg of vector.

2.8.2 Ligation and transformation

Pvu II-cut and phosphatased pAT 153/Pvu II/8 or pATX vectors were kindly provided by K. T. Belt and L.C. Wu respectively. The bulk or gel-eluted restriction endonuclease digested, blunt-end DNA fragments (2-200 ng) were mixed with 2-10 ng of plasmid vector, in 1 mM rATP/50 mM Tris pH 7.4/10 mM MgCl_2 /20 mM DTT/50 $\mu\text{g/ml}$ BSA/0.2 mM spermidine and 2 U of T4 DNA ligase in a final volume of 10 μl . This was incubated at room temperature for 5 hr. 100 μl of competent cells was added to the ligation mixture, left on ice for 10 min and 37°C for 5 min before addition of 4 volumes of L broth. The transformed cells were incubated at 37°C for 1 hr to allow expression of antibiotic marker. 50 and 200 μl of transformed cells were plated onto L agar/ampicillin (LA) plates and incubated overnight at 37°C . 50-1000 bacterial colonies were spotted onto duplicate LA plates with grids and grown for 5-8 hr. Bacterial colonies on one of the duplicates were transferred onto Whatman 541 filters. The filters and the plates were also marked in an identical fashion so they can be matched up later. The other (master) plates were sealed with parafilm and stored at 4°C while filters were processed.

2.8.3 Screening of colonies for positives (Denhardt, 1966; Gergen et al., 1979)

The colonies were lysed in situ by floating the filters with colony side up on the following solutions

- a) 0.5 M NaOH 5 min
- b) 0.5 M Tris, pH 7.4 5 min

c) 0.5 M Tris, pH 7.4 5 min

(1 x SSC is 150 mM NaCl/15 mM Na Citrate, pH 7.0)

and followed by brief dipping into 95% ethanol. Filters were dried on Whatman 3 MM paper at room temperature.

The filters were prehybridised at 42°C for at least 2 hr in 5 x SSC/0.2% Ficoll/0.02% BSA/0.02% polyvinylpyrrolidone/10% dextran sulphate 50% formamide, containing 100 µg/ml sheared salmon sperm DNA and 10 µg/ml *E. coli* with pAT 153/Pvu II/8 DNA which had been boiled for 5 min and 'snap-cooled' before use. The filters were hybridised for 16 hr in the same buffer containing $\geq 2 \times 10^5$ dpm/ml of [³²P]-labelled probe, also denatured by boiling before use.

After hybridisation the filters were washed as follows:

4 times in 2 x SSC/0.1% SDS for 15 min at room temperature.

2 times in 0.1 x SSC/0.1% SDS for 30 min at room temperature.

The filters were air dried and autoradiographed. The autoradiograph was used to match with the original plates via the markings on the filters. Positives were picked from the other unprocessed master plate.

2.9 Organic synthesis of oligonucleotides by the phosphotriester method. (Sproat and Gait, 1984).

A 20-mer oligonucleotide, (5' CAGGAGACATCTAACTEECT 3') which corresponds to the DNA sequence 30 nucleotides 5' to the C4d isotypic region (Yu et al., 1986) was chemically synthesized and used as DNA sequencing primer (Section 2.10).

The phosphotriester solid-phase synthesis method was used. In this method a 5'-O-protected deoxyribonucleotide corresponding to the first 3' nucleotide of the primer was attached to a solid support. Chain assembly was by alternating terminal 5'-deprotection reactions and

coupling reactions. In both cases excess reagents were added to drive the reaction to completion and unreacted components removed by washing of the support with an appropriate solvent. Cycles of synthesis were continued until the required length was obtained and then the oligonucleotide was cleaved from the support, protecting groups removed, and the deprotected oligonucleotide purified.

2.9.1 Support functionalisation and chain assembly

(a) support functionalisation.

The oligonucleotide was synthesized by a semi-manual Cruachem module. Two micromole of resin, polydimethylacrylamide-kieselguhr thymine deoxyribonucleoside, was packed in a reaction column. Air bubbles were removed by gentle stirring of the resin in pyridine. The resin was treated with 10% V/V phenylisocyanate: pyridine for 10 min. The resin was functionalised as the 5' protecting group on the nucleotide was removed.

(b) coupling

The column was washed with pyridine. As oligodeoxyribonucleotides are synthesized from the 3' end to the 5' end, dextrodeoxynucleotide phosphodiester triethylammonium salt which correspond to the second most 3' residue (i.e. cytosine) was added. One-shot pack of such nucleotide was dissolved in 230 μ l pyridine, of which 115 μ l was added to a glass vial containing 25 mg coupling reagent, 1-mesitylene-sulphonyl-3-nitro-1,2,4-triazole (MSNT). 10 μ l of catalyst, 1-methylimidazole, was added. The reagents were injected to the reaction column. After 15 min, excess reagents were removed by pyridine wash, subsequently, dichloroethane was added to the column.

(c) deprotection

The 5' protecting group of the nucleotide on the resin was removed by washing with 3% (v/v) dichloroacetic acid in dichloroethane, 40 sec for nucleotide A or G, or 75 sec for nucleotide C or T. Afterwards, the column was washed with excess dichloroethane, and then with pyridine.

(d) The coupling, and deprotection cycles (b and c) were repeated for other nucleotides until the last 5'-nucleotide was coupled.

2.9.2 Deprotection and purification of the oligonucleotides

(a) Detachment of oligonucleotides from resin and removal of 2-chlorophenyl groups from the phosphate.

After the last coupling reaction, the column was washed with pyridine. The resin was transferred to an Eppendorf vial and washed with 1 ml 50% (v/v) dioxan in water, twice. 150 mg 2-pyridinealdoxime, 125 μ l tetramethylguanidine in 1 ml 50% dioxane was added and reaction mixture was left in dark for 16 hrs, at room temperature. The 2-chlorophenyl protecting groups were removed and the succinate linkage was cleaved. The supernatant was transferred to a round-bottomed flask. The resin was washed with 50% dioxan and the supernatant was pooled, evaporated using a rotatory evaporator, under reduced pressure and the use of a 60°C water bath.

(b) Acyl groups and 5' protection groups

To the flask with processed oligonucleotides, 5 ml concentrated ammonia solution (about 35% ammonia by weight, specific gravity 0.88) was added, the flask was sealed and left at 50°C for 6 hrs. The acyl protecting groups were removed. The flask was cooled down at room temperature and left unstoppered in the fume-hood overnight. The sample was then dried using a rotatory evaporator as described. To the dried sample, 5 ml water was added and the evaporated, to remove residual

ammonia. The evaporated residue was treated with 5 ml 80% (v/v) acetic acid in water for 30 min to remove the 5'-terminal protecting group. The sample was evaporated and subsequently dissolved in 5 ml water.

(c) Purification of oligonucleotides

Organic compounds were removed by either extraction, twice. Residual ether was removed by evaporation and the dried sample resuspended in 4 ml of water. The suspension was desalted on a G-25 column, using 20% ethyl alcohol as eluant. 3 ml fraction were collected and oligonucleotide concentration was monitored by a UV-spectrometer. Fraction peaks were pooled and dried. The oligonucleotides were further purified by gel electrophoresis, using a 20%, denatured acrylamide gel. After electrophoresis, oligonucleotide band corresponding to the 20-mer was excised, as visualized under a fluorescent thin layer chromatography plate and overhead UV source. The oligonucleotide in the gel was eluted with 1 ml water in a 37°C shaken for 16 hr. The eluant was desalted in a G25 column to remove urea and fraction-collected as described. Breakthrough peak under UV was pooled and dried. The purified oligonucleotides was resuspended in 50 µl of 10 mM Tris, pH 7.4, 0.1 mM EDTA and kept in -70°C.

2.10 Dideoxy DNA sequencing

2.10.1 Preparation of vector (Messing, 1983)

10 µg of the replicative form of bacteriophage M13 mp8 or mp9 was cleaved with 30 units of Sma I enzyme at 25°C for 3-4 hr; then with 5 µg of calf intestinal alkaline phosphatase at 37°C for 30 min. The DNA was extracted extensively with phenol/chloroform before recovery by ethanol precipitation. Sometimes it was necessary to repeat the phosphatase reaction.

2.10.2 Sub-cloning of DNA fragment onto M13 vectors

a) Random (Deininger, 1983)

Approximately 10 µg restricted DNA was isolated from an agarose gel (Section 2.7.2). The sample was sonicated in a MSE sonicator for 30 sec (i.e. 6 x 5 sec) on ice, breaking the DNA at random positions. The sheared fragments were repaired in the presence of all four nucleotides (Section 2.5.1) using E. coli polymerase I Klenow fragment. DNA fragments between 300-1000 base pairs in size were recovered by a 1.5% agarose gel.

b) Non-random

DNA restriction fragments were isolated from acrylamide or agarose gels as described in Section 2.7.

c) Ligation

20 ng cut and phosphatased vector was mixed with 5-50 ng DNA in 1 mM rATP/50 mM Tris, pH 7.4/10 mM MgCl₂/20 mM DTT/50 µg/ml BSA/2 mM spermidine in a final volume of 10 µl. About 2 units of T4 DNA ligase was added and the ligation reaction was carried out at 16°C for 16-24 hr (Winter et al., 1981).

d) Transformation of E. coli JM 103 competent cells

The cells were made competent by the following treatment. 500 µl of an overnight culture of JM 103 was diluted into 50 ml 2 x TY broth and grown to an A₆₀₀ of 0.4-0.6 OD units. The cells were pelleted by centrifugation at 2000 g at 4°C for 10 min in the MSE-6L centrifuge. The cells were resuspended in 10 ml 50 mM CaCl₂/10 mM Tris, pH 7.4 and left on ice for 30 min. The cells were pelleted again, resuspended in 6 ml of the same buffer and stored at 4°C. The cells were normally competent for up to 3 days.

200 µl competent cells were added to the ligation mix, incubated at 0°C for 45 min and then heat-shocked at 42°C for 5 min.

The transfected cells were added to 3 ml molten L top agar containing 25 μ l of 2.5% aqueous solution of isopropylthiogalactoside (IPTG) and 50 μ l of 2% bromochlorindolylgalactoside (BCIG in dimethylformamide, kept at 45°C. Finally, 0.5 ml of overnight JM 103 cells was added, gently mixed and poured onto 2 x TY agar plates.

Incubation of the plates overnight at 37°C resulted in the production of blue and white plaques; areas of lower density cells caused by M13, a non-lytic persistent phage, reducing the growth rate of infected cells. Blue plaques represent bacteria containing M13 which has β -galactosidase activity due to complementation between the M13 and JM 103. IPTG induces enzyme synthesis which hydrolyses BCIG, a substrate analogue, releasing a blue indolyl derivative. The cloning site in M13 is within the β -galactosidase gene. The M13 can no longer complement the 'M 15' protein of the host (Messing et al., 1977), no dye is released and white plaques will result.

2.10.3 Preparation of phage DNA (Sanger et al., 1980)

Agar plugs of white plaques were grown in 1.5 ml of one hundredth dilution (15-18 μ l) of an overnight culture of JM 103 in 2 x TY broth at 37°C for 5 hr, with vigorous, constant shaking. The medium was transferred to an Eppendorf tube and spun for 5 min in a microfuge at room temperature. The supernatant was transferred to a fresh tube and respun for 5 min. Approximately 1.2 ml of this supernatant was transferred to another tube containing 200 μ l 20% polyethylene glycol/2.5 M NaCl, mixed and kept on ice for 30 min. This was centrifuged for 10 min, the supernatant discarded, and all residual PEG removed. The pellets were resuspended in 100 μ l 10 mM Tris/0.1 mM EDTA pH 8.0, phenol/chloroform

extracted and precipitated with 10 μ l 3 M NaAc pH 6.0 and 300 μ l 95% ethanol at -70°C for 30 min. The DNA was pelleted, washed in 70% ethanol, dried and resuspended in 30 μ l or 50 μ l Tris/0.1 mM EDTA, pH 8.0. The single strain phage DNA was stored at -20°C .

2.10.4 Relative orientation of integrated fragments (Herrmann et al., 1980)

The orientation of the cloned restriction fragments was determined as follows. 2 μ l of each clone was mixed with 2 μ l of a reference clone in a sealed glass capillary in the presence of 4 μ l 600 mM NaCl/70 mM MgCl_2 /60 mM Tris, pH 7.4 and 1 μ l 50% glycerol/1% SDS/bromophenol blue. The tubes were incubated at 67°C for 30-60 min, snap cooled in icy water and electrophoresed on a 0.8% agarose minigel. Cloned DNA containing fragments of opposite orientation hybridised with each other and migrated slower due to the increase in size.

2.10.5 Sequencing reaction (Sanger et al., 1977; Biggin et al., 1983; Williams et al., 1986)

Besides the application of a chemically synthesised oligonucleotide for sequencing the C4d isotypic region (Section 2.9), a universal oligonucleotide primer (Duckworth et al., 1981) used to prime for synthesis using single-stranded DNA as template for all sequencing reactions. The sequence of the oligomer was:

5'd (GTAAAACGACGGCCAGT) 3'

The composition of the mixes used was:

'A' mix :- 0.125 mM dCTP; 0.125 mM dGTP; 0.125 mM dTTP: 15 μ M ddATP

'C' mix :- 6.25 μ M dCTP; 0.125 mM dGTP; 0.125 mM dTTP: 40 μ M ddCTP

'G' mix :- 0.125 mM dCTP; 6.25 μ M dGTP; 0.125 mM dTTP: 80 μ M ddGTP

'T' mix :- 0.125 mM dCTP; 0.125 mM dGTP; 6.25 μ M dTTP: 250 μ M ddTTP

a) Sequencing using ^{32}P -dATP

Approximately 2ng primer was hybridised with 8 μl single-stranded DNA (initial volume = 50 μl) in 13.6 mM Tris/6.8 mM MgCl_2 , pH 8.5, in a total volume of 11 μl , at 60°C for 1 hr. 2 μl of primed clone were added to four tubes: T, C, G, A. To this was added 2 μl of the correct mix of 2 μl of enzyme mix: 2 U DNA polymerase I, 'Klenow' fragment (Boehringer or Pharmacia) and 2.5 μCi [α - ^{32}P]-dATP in 11 mM DTT/0.15 μM dATP/10 mM Tris pH 8.0.

The reaction was started by mixing the reactants in a capless Eppendorf tube in a microfuge. The reaction was incubated at room temperature for 15 minutes and chased for another 15 minutes with 2 μl 0.5 mM dNTPs. The reaction was stopped by the addition of 4 μl formamide dye mix (Section 2.6.1), and the samples boiled for 3-4 min and snap-cooled.

b) Sequencing using ^{35}S dATP

The sequencing reaction was virtually identical to ^{32}P -dATP sequencing method except with the following modifications:

- i) M13 DNA (Section 2.10.3) was resuspended in 30 μl instead of 50 μl so as to increase the M13 template concentration.
- ii) The first sequencing reaction (i.e. prior to the chase reaction) was left for 20 min instead of 15 min.
- iii) About 1.5-2.0 μl of reaction mixture was loaded to the gel instead of 0.5-1.0 μl .

2.10.6 Electrophoresis of sequenced samples

The samples were run on either (a) 6% denaturing gels (Section 2.6.1) at 25 mA for 4-5 hr or (b) buffer gradient gels at 30 mA for 1 hr 20 min (Biggin et al., 1983). Prior to preparing both types of gel the back plate was treated with 2.5 ml ethanol containing 75 μl 10% acetic

acid and 7.5 µl 'Wackersilicone GF 31' (Wacker Chemie, W. Germany) (Garoff and Ansorge, 1981). The solution was rubbed well onto the glass then washed with ethanol and thoroughly wiped to removed excess silane. The silane binds the gel covalently to the back plate, whereas the solution used on the notched plate repels the gel.

After electrophoresis the gels adhering to the silane treated plates were fixed in 10% acetic acid/10% methanol for 15 min and then washed in water for 15 min. The urea free gels were dried for about 30 min at 100-120°C, and then autoradiographed for 1-3 days at room temperature, without a screen, and the film was not pre-flashed.

Buffer gradient gels

0.5 TBE gel mix :- 6% acrylamide/0.5 x TBE/7 M urea

2.5 TBE gel mix :- 6% acrylamide/2.5 x TBE/7 M urea/5%
sucrose/0.05% bromophenol blue

7 ml 2.5 TBE mix were mixed with 35 µl 10% ammonium persulphate and 14 µl TEMED. 35 ml 0.5 TBE mix were mixed with 175 µl 10% ammonium persulphate and 70 µl TEMED. 6 ml 0.5 TBE mix were taken up into a 10 ml pipette, followed by 6 ml 2.5 TBE mix, and a gradient formed by introducing 5 air bubbles through the interface. This mix was poured between the gel plates and the remaining space filled with the 0.5 TBE mix. 0.5 x TBE was used as the running buffer in the top reservoir and 1 x TBE in the bottom.

2.10.7 Analysis of nucleotide sequence data (Staden, 1982a,b,1984)

Sequence data from the random cloning were analysed using a Vax II/780 computer. Sequences were read from the gels and entered into the computer by hand using BATIN. They were screened for M13 and pAT153 sequences using SCREEN V. Non-vector sequences were overlapped and

aligned using DBAUTO and DBUTIL. Other programmes used in data analysis include DBSTART, DAIGON, ANALYSEQ, ANALYSEP and HYDROPLOT.

2.11 DNA transfer blot analyses (Southern, 1975; 1979; Wahl et al., 1979)

2.11.1 Analysis of DNA fragments between 0.5-15 kb in size

- a) Gel electrophoresis, treatment and capillary blotting of DNA to nitrocellulose filters.

Samples in 1/5 volume of glycerol dye mix were electrophoresed in 0.8% agarose gels/1 x TBE (Section 2.6.3) at 50 mA for approximately 16 hr. The gel was stained with ethidium bromide, photographed and then washed with the following solutions:

- i) Twice in 0.25 M HCl for 15 min at room temperature.
- ii) Twice in 0.5 M NaOH/1 M NaCl for 15 min at room temperature.
- iii) Twice in 3 M NaCl/0.5 Tris, pH 7.4 for 15 min at room temperature.

The gel was placed on two sponges and a piece of Whatman 3 mm paper soaked in 20 x SSC. The surrounds of the gel were covered with cling-film and glass strips placed at the ends of the gel. A piece of nitrocellulose (Amersham Hybond C) briefly floated on water was placed on top of the gel, and this was covered with two sheets of dry Whatman 3 mm paper, approximately three inches of paper towels weighted with a glass plate and a book. The buffer passed through the gel and nitrocellulose carrying the DNA from the agarose to the nitrocellulose.

With clones DNA blots were taken for 30 min to 2 hr. With genomic DNA blotting was for 24 hr. The sponges were kept moist with additional 20 x SSC and the paper towels were replaced several times. The DNA was baked onto the nitrocellulose for 2-3 hr at 80°C before hybridisation.

b) Hybridisation of nitrocellulose filters (Bernards & Flavell, 1980)

Filters were washed in 1 M NaCl/1 mM EDTA/50 mM Tris, pH 8.0/0.1% SDS for 1 hr at 42°C. Prehybridisation was carried out at 42°C for at least 3 hr in 50% deionised formamide/1 M NaCl/50 mM Tris, pH 7.4/10% dextran sulphate/0.2% Ficoll/0.2% BSA/0.2% polyvinylpyrrolidone/0.1% sodium pyrophosphate/0.1% SDS, (hybridisation buffer) containing 100 µg/ml sonicated salmon sperm DNA and 10 µg/ml *E. coli*/pAT 153 Pvu II-8 DNA. This mixture was boiled for 5 min and cooled before use. The filters were hybridised for 48 hr in the above buffer containing $\sim 5 \times 10^5$ dpm/ml of [^{32}P]-labelled probe, also boiled before use.

After hybridisation the filters were washed as follows:

3 times in 2 x SSC/0.1% SDS for 10 min at room temperature

2 times in 0.1 x SSC/0.1% for 30 min at 68°C.

The filters were air dried and autoradiographed (Section 2.9.3).

2.11.2 Analysis of DNA fragments between 0.1-1.5 kb in size

a) Gel electrophoresis, treatment and capillary blotting of DNA onto nylon membrane

Samples in 1/5 volume of glycerol dye mix were electrophoresed in 1.5% agarose gel (Pharmacia)/1 x TBE (Section 2.6.3) at 50 mA for approximately 16 hr. The gel was stained with ethidium bromide, photographed and then washed with the following solutions with gentle agitation:

i) Twice with 0.4 M NaOH/0.6 M NaCl at room temperature for 15 min.

ii) Twice with 1.5 M NaCl/0.5 M Tris pH 7.5 at room temperature for 15 min.

The gel was placed on two sponges and a piece of Whatman 3 mm paper soaked in 20 x SSC. The surrounds of the gel were covered with cling film as described. A piece of Gene Screen Plus membrane had been wetted with distilled water and soaked in 20 x SSC for 15 min was placed on the top of the gel. Uncovered agarose gel was trimmed.

The concave side (when the membrane was dry) of the membrane was in contact with the gel. Two sheets of Whatman paper, paper towels and small weight on top of the towel were applied as described in the previous section (2.11.1). The DNA was capillary transferred onto the nylon membrane through 24 hr blotting. After blotting, immobilised DNA on the nylon membrane was further denatured completely by immersing membrane in 0.4 M NaOH for 60 sec and then neutralized in an excess of 0.2 M Tris pH 7.5/2 x SSC. The membrane was then dried at room temperature without further baking.

b) Hybridisation of Gene Screen Plus membrane

The membrane was prehybridised, hybridised and processed as described in 2.11.1.b except that the hybridisation buffer and 2 x SSC were enriched with 1% SDS instead of 0.1%, so as to reduce background noise after hybridisation.

2.12 Construction of genomic libraries (Maniatis et al., 1982; Frischauf et al., 1983; Kaiser and Murray, 1985)

2.12.1 Preparation of the DNA for cloning

Genomic DNA was isolated from white blood cells (Section 2.3.6). The DNA was digested in 4 separate reactions (50 µg DNA per reaction) at 37°C for 1 hr, with Mbo I at concentrations of 0.3 U/µg, 0.13 U/µg, 0.07 U/µg, and 0.02 U/µg. the reactions were terminated with 10 mM EDTA and heating to 70°C for 15 min. The DNA was extracted with equal volumes of

phenol; phenol : chloroform : isopropanol (25:24:1); and chloroform : isopropanol (24:1). Each time the aqueous layer was transferred to a fresh tube and the organic layer re-extracted with 10 mM Tris/1 mM EDTA, pH 8.0 to minimise DNA loss. The DNA was then ethanol precipitated, washed, dried and resuspended in 100 μ l 10 mM Tris/1 mM EDTA, pH 8.0.

The DNA was fractionated on a 38 ml 10 - 40% sucrose gradient in the presence of 1 M NaCl/20 mM Tris, pH 8.0/10 mM EDTA. The gradient was centrifuged in the Beckman L5-65 ultracentrifuge using the SW-27 rotor at 26,000 r.p.m. and 20°C for 20 hr. Fifty 10 drop fractions (0.7 ml) were collected and 5 μ l aliquots analysed on a 0.4% agarose gel. Fractions containing DNA in the size ranges 15-20 kb, 20-30 kb and above 30 kb were pooled, ethanol precipitated twice, washed, dried, resuspended in 10 mM Tris/1 mM EDTA, pH 8.0 and stored at 4°C. The DNA concentration was estimated by spotting 1 μ l onto 1% agarose plates in 10 mM Tris/1 mM EDTA, pH 7.4/1 μ g/ml ethidium bromide. λ DNA of known concentrations were used as standards and the plates were photographed over an ultra-violet light transilluminator (Section 2.6.3).

2.12.2 Preparation of the λ vector arms

20 μ g of CsCl purified λ -EMBL-3 DNA were digested with excess Bam HI enzyme at 37°C for 2 hr. The reaction was terminated with 15 mM EDTA and heated to 70°C for 10 min. The sample was ethanol precipitated, washed and the DNA resuspended in 10 mM Tris/0.1 mM EDTA, pH 7.5. The DNA was then digested with excess EcoR I enzyme at 37°C for 10 min. The DNA was extracted with equal volumes of phenol: chloroform : isopropanol (25:24:1) and chloroform : isopropanol (24:1). DNA in aqueous phase was differentially precipitated with 0.15 volumes of 0.3 M NaAc, pH 6.0/0.6 volumes isopropanol, after incubation on ice for 15 min. DNA pellet was washed with 500 μ l of 0.35 M NaAc pH 6.0: ethanol (1.25 v/v). Thus

small linker fragments generated by BamH I-EcoR I digests was removed. A second normal ethanol precipitation was followed and the dried vector DNA was resuspended in 10 mM Tris pH 8.0/0.1 mM EDTA at 0.5 µg/µl and stored at -20°C.

2.12.3 Packaging extracts and packaging reaction

Packaging extracts, freeze-thaw lysate and sonicated extract (isolated from E. coli BHB 2688 and BHB 2690 respectively) were kindly provided by K. T. Belt. Efficiency of the packaging extract was tested by mixing 0.1 µg of uncut λ EMBL-3 DNA, 5 µl of buffer A (20 mM Tris pH 8.0/5 mM MgCl₂/0.5% (v/v) β-mercaptoethanol/1 mM EDTA), 1 µl buffer M1 (6 mM Tris, pH 7.5/15 mM spermidine/30 mM putrecine/18 mM MgCl₂/0.2% rATP/30 mM β-mercaptoethanol), 3.5 µl sonicated extract and 5 µl freeze-thaw lysate. After incubating at 25°C for 1½ hr 230 µl phage buffer (6 mM Tris, pH 7.5/10 mM MgCl₂/100 mM NaCl/0.5 mg/ml gelatin) was added. Packaged DNA (i.e. phage) were serially diluted and plated on E. coli Q 358. Packaging efficiency was about 1×10^7 phage particles per µg of λ EMBL-3 DNA.

2.12.4 Ligation, packaging and assay of recombinants

0.5-1.5 µg of vector arms were mixed with 0.5-1 µg of size fractionated genomic DNA (15-20 kb) size range) in 1 mM rATP/50 mM Tris pH 7.4/10 mM MgCl₂/20 mM DTT/50 µg/ml BSA and incubated with 2 U DNA ligase at 16°C for 20 hr, in a total volume of 10 µl.

3 µl of ligated DNA extract was packaged as described in 2.12.3. The package recombinants were stored at 4°C with an aliquot assayed by plating on E. coli Q 359. To assay the titer of recombinants 1 µl of diluted, packaged phage was mixed with 0.5 ml of overnight culture of E. coli Q 359 ($A_{600} \sim 1.2$ units) which was grown in T broth/0.2% maltose,

and enriched with 10 mM MgCl_2 just before infection of phage. Infection of phage was carried out at 37°C for 30 min. 3.0 ml of molten T agar was added and the mixture plated on T bottom agar plate. Very tiny recombinant phages were counted after 24 hr incubation at 37°C.

2.12.5 Plating and screening of λ EMBL-3 genomic library

(a) Plating

About $1-2 \times 10^6$ recombinants from various ligation tests were pooled and split to aliquots with each about 10^5 phage particles. To each aliquot 2 ml of overnight culture Q 359 (prepared as described) was added and the bacteria was infected at 37°C for 30 min. The mixture was added by 9 ml of T top agarose (held at 48°C), poured onto 14 cm T agar plates, and incubated at 37°C overnight.

(b) Screening

After the overnight incubation at 37°C plates were left at 4°C for at least one hour to allow the top agarose to harden. Dry nitrocellulose filters (millipore) were placed in direct contact with the plaques and keyed with ink asymmetrically, marking both filter and solid medium. After 60 sec the filter was removed and immersed, DNA side up, in the following solutions:

1.5 M NaCl/0.5 M NaOH (denaturing solution) for 60 sec.

1.5 M NaCl/0.5 M Tris, pH 7.5 (neutralising solution) for 5 min,

2 x SSC for 5 min

and then air dried on Whatman 3 mm paper. Two replicas were taken from each plate, using the original ink marks on the plate for the subsequent copy.

The DNA was baked onto the nitrocellulose at 80°C for 2 hr before hybridising, washing and autoradiography (Section 2.11.1).

(c) Re-screening

Agar plugs were taken from the original plates into 0.5 ml phage buffer/1% chloroform. E. coli Q 358 cells in T broth/0.2% maltose were grown at 37°C for 20 hr when the A_{600} was ~ 1.8 units. 10^2 - 10^3 phage were added to 0.5 ml Q 358 cells in 10 mM $MgCl_2$ and incubated at 37°C for 30 min. The mixture was added to 3 ml T top agarose, mixed, poured onto 9 cm T agar plates and incubated at 37°C overnight. The plaques were re-screened using the method described above. Filters screened with oligonucleotides were pre-hybridised for 3 hr in hybridisation buffer (Section 2.11.2) without formamide and hybridised overnight with $\sim 10^6$ cpm/ml [$\gamma^{32}P$]-labelled oligonucleotide, pre-boiled for 5 min before use. The hybridisation temperature was 2°C below the melting temperature of the oligonucleotide. After hybridisation the filters were washed as follows:

twice in 6 x SSC for 15 min at room temperature

twice in 6 x SSC/0.1% SDS for 15 min at the hybridisation temperature.

The filters were then air dried and autoradiographed.

Single isolated plaques were chosen from the re-screen plates and used to prepare stocks of infectious particles, stored at 4°C in phage buffer/0.3% chloroform. These were used to give plate lysates from which DNA was isolated (Section 2.3.4) for analysis by restriction mapping and nucleotide sequencing.

CHAPTER THREE

Molecular structure of a human complement component C4A gene

3.1 Introduction

Between 1981 - 83, partial protein sequence around the thiolester site of complement C4 was obtained (Campbell et al., 1981; Harrison et al., 1981; Chakravarti et al., 1983). This protein sequence enabled the deduction and synthesis of a C4d mixed oligonucleotide probe which was used to isolate cDNA clones. The first clone, Alu-7, contained a 301 bp insert corresponding to the C4d region, was screened out from a human liver cDNA fragment library (Carroll and Porter, 1983). This small cDNA was used to isolate full length cDNA clones, which were subsequently used to probe genomic cosmid libraries (Belt et al., 1984; Carroll et al., 1984a; b).

The complete cDNA sequence of a C4A allele, pAT-A, and an almost complete cDNA sequence of a C4B allele, pAT-F, were obtained. The derived amino acid sequence from pAT-A suggested that the single chain precursor molecule of C4, i.e. pro-C4, contains 1722 amino acid residues, although sequence data from pAT-F inferred that pAT-A may contain a 3-residue deletion near the end of the alpha chain. The size of the beta, alpha and gamma chains of secreted C4 is estimated to be 656, 767 and 291 amino acid residues respectively. The cDNA sequences also suggested that C4 contains four potential N-linked glycosylation sites; one on the beta chain at residue 207, three on the alpha chain at residues 843, 1309 and 1372, and none on the gamma chain.

The availability of genomic cosmid clones has enabled the

genomic sequence of the complement C4 to be determined. This is crucial and fundamental in the following aspects:

a. Determination of exon-intron structure of the C4 gene.

As described in section 1.3, mammalian genes are split into coding sequences (exons) separated by intervening sequences (introns) which are spliced out in the mature RNA transcript (Gilbert, 1977; 1985). There seems to be some correlation between the exons and the discrete regions of expressed protein molecule. They often code for useful portions of protein structure such as functional regions, folding elements, domains or subdomains. Moreover, exons appear to be independent segments of DNA which can be sorted independently (Gilbert, 1985). Therefore, determining the exon/intron structure of C4 may give insights to the protein structure and function of C4, particularly if the 3-dimensional structure of the protein is also elucidated. It would also give clues to the evolutionary history of the C4 gene and its relationship with other related proteins such as complement components C3, C5 and the protease inhibitor, α_2 macroglobulin (α_2M), especially if the organisation of these genes is also known.

b. Regulation of C4 gene expression

Genomic sequences at the 5' and 3' flanking regions are involved in the regulation of gene expression, especially for transcription initiation and termination. Thus determination of the C4 flanking gene sequences will facilitate elucidation of cis-acting DNA elements such as enhancer(s), promoter and other

motifs which may be involved in the cell-specific, and probably inducible expression of C4.

c. Polymorphism of C4

C4 is one of the most polymorphic complement proteins. C4 polymorphism can be detected at the genomic level, and knowledge of the C4 genomic sequence will facilitate these studies. For instance, information about the exact location of polymorphic sites, and the presence of nearby restriction enzyme cleavage sequences can be obtained. These will be useful for the generation of specific DNA probes and for the identification of informative restriction fragment length polymorphisms (RFLPs). By providing background information, it may also facilitate further studies on different rare alleles such as C4A 1 and C4B 5, which express essentially the reversed Rodgers (Rg) / Chido (Ch) antigenicity (Rittner et al., 1984; Roos et al., 1984); and the C4A 6 allele, which has a lower hemolytic reactivity compared with other C4A allotypes (Dodds et al., 1985).

A cosmid clone of C4, Cos 3A3, was used for the determination of the C4 gene sequence. Cos 3A3 contains a DNA insert ~40 kb in the pTCF vector (Carroll et al., 1984a; Grosveld et al., 1982), of which ~20 kb encodes complement component C4. DNA fragments encoding C4 were subcloned into plasmid vector pAT 153/Pvu II/8. From these plasmid subclones, more accurate restriction maps were constructed and appropriate DNA fragments were isolated and sequenced. DNA sequence was typed into the VAX computer and aligned by the computer programs, BATIN, DBAUTO and DBUTIL (Staden 1982^{a,b}). The exon/intron structure of C4 was deduced by comparing the genomic DNA sequence with the published cDNA sequence (Belt et al., 1984).

As explained in chapter 4 and by Belt et al., (1985), the C4 gene in Cos 3A3 is probably a C4A 3 allele (sub-type a).

3.2 Subcloning, restriction mapping and DNA sequencing

Restriction sites for Sal I, Hind III and Cla I of Cos 3A3 had been mapped by Mike Carroll (Carroll et al., 1984a; Fig. 3.1). Southern blot analysis using full length cDNA probe, pAT-A, suggested that the coding region of the C4 gene lies within the (I) 8.0 kb Hind III-Cla I, (II) 6.7 kb Cla I-Hind III and the (III) 9.9 kb Hind III-Sal I restriction fragments. Within the 8.0 kb and the 6.7 kb fragments there is a big intron of about 6-7 kb. These three restriction fragments were blunt-end ligated into the pAT-153 vector and the corresponding subclones designated pCHS1-1, pCHS2-1 and pCHS2-10 respectively. More detailed restriction enzyme maps were constructed for each subclone as shown in Figs 3.2, 3.3 and 3.4. In order to generate smaller DNA fragments to facilitate mapping and DNA sequencing, the following smaller subclones were constructed:

<u>Subclone</u>	<u>Size (kb)</u>	<u>Digest</u>	<u>Parent plasmid</u> <u>or cosmid</u>
1. pBH4	4.1	Hind III - Bgl II	Cos 3A3
2. pRP-1	1.3	Bgl II - Bgl II	pCHS2-1
3. pAB-3	2.1	Acc I - Bst EII	pCHS1-1
4. pAB-B7	3.4	Bst EII - Sal I	pCHS1-1

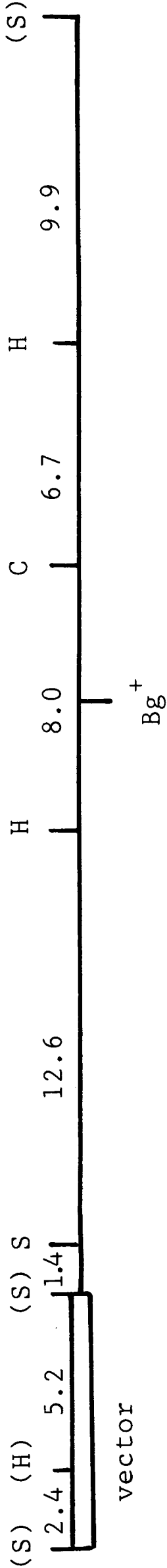
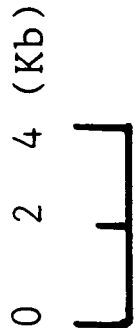
Southern blot analysis of the subcloned DNA established that the boundaries of the 6-7 kb intron are within the 1.3 kb Bgl II fragment of pCHS2-1 and the 450 bp Stu I fragment of pCHS2-10 (Figs. 3.2 and 3.3). Based on the restriction mapping of the

Legend to Fig. 3.1

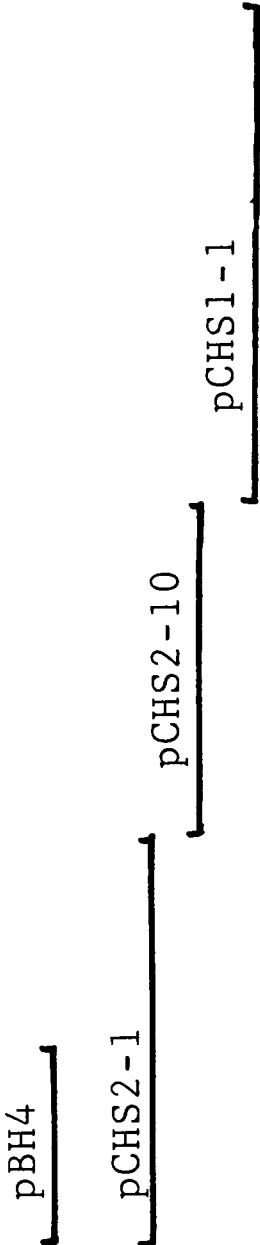
Published restriction map of Cos 3A3 (Carroll et al., 1984a)

Cos 3A3 contains a ~40 kb insert cloned in the Bam HI site of the pTCF vector (Grosveld et al., 1982). Southern blot analysis using full length cDNA probe of C4 suggested that the C4 gene is located within the 8.0 kb Hind III - Cla I, 6.7 kb Cla I - Hind III and the 9.9 kb Hind III - Sal I restriction fragments. These fragments were subcloned into pAT153 /Pvu II/8 vector by blunt-end ligation (see Material and Methods). The corresponding subclones were designated as pCHS2-1, pCHS2-10, pCHS1-1, respectively. An additional subclone of the 4.0 kb Hind III - Bgl II fragment at the 5' region had been constructed and designated as pBH4. Restriction sites of the pTCF vector are enclosed by brackets. Abbreviations: Bg⁺, Bgl II; C, Cla I; H, Hind III; S, Sal I.

Restriction map of Cos 3A3



subclones :

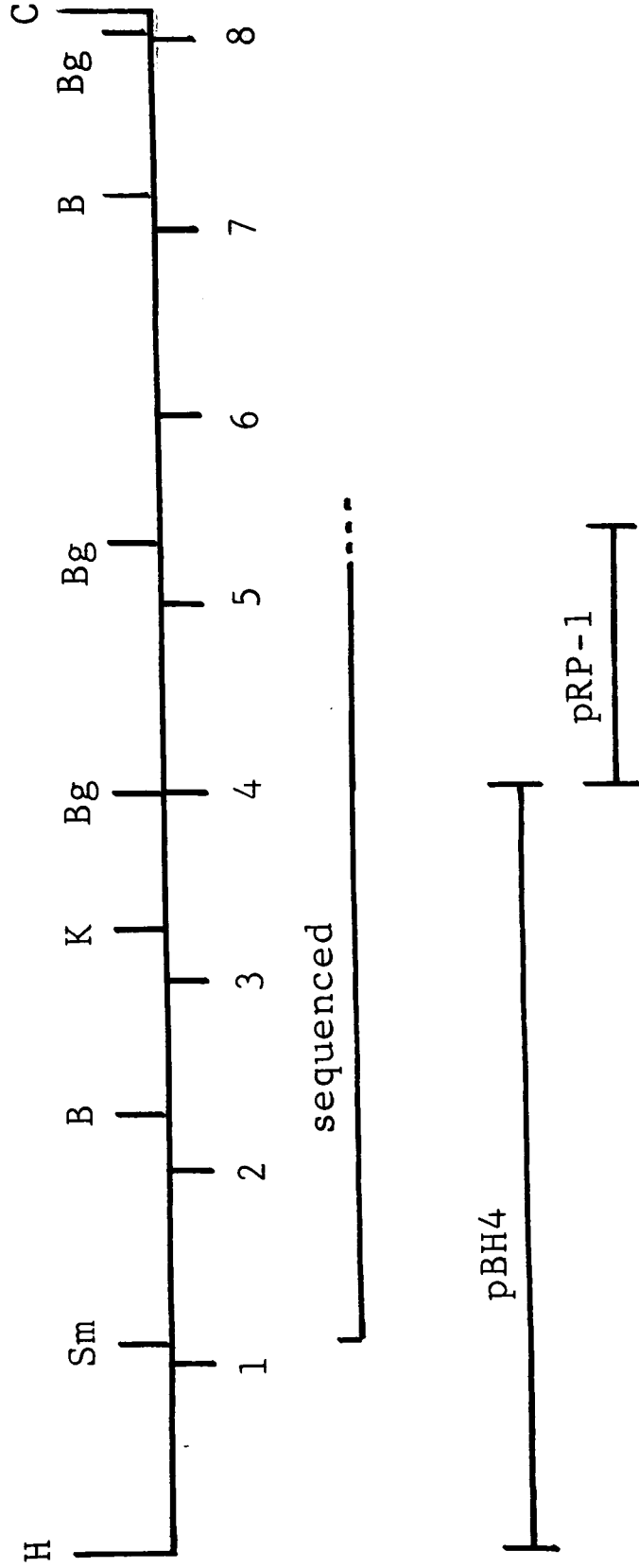
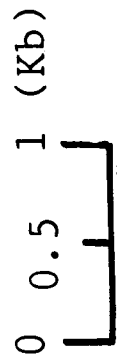


Legend to Fig. 3.2

Restriction map of pCHS2-1

Plasmid DNA was used for various single and double restriction digests. The location and size of restriction fragments were deduced from Southern blot analysis, using a 5' specific cDNA probe (P_A), which is a 476 bp BamH I - Kpn I restriction fragment of pAT-A, and a full length cDNA probe from pAT-A. Most of the restriction sites were confirmed by DNA sequencing. pRP-1 is a subclone containing the 1.3 kb Bgl II restriction fragment as shown. The region where DNA sequence was determined is indicated by a solid line. Partially elucidated DNA sequence is represented by a dotted line. Abbreviations: B, BamH I; Bg, Bgl II; C, Cla I; K, Kpn I; Sm, Sma I.

Restriction map of CHS2-1



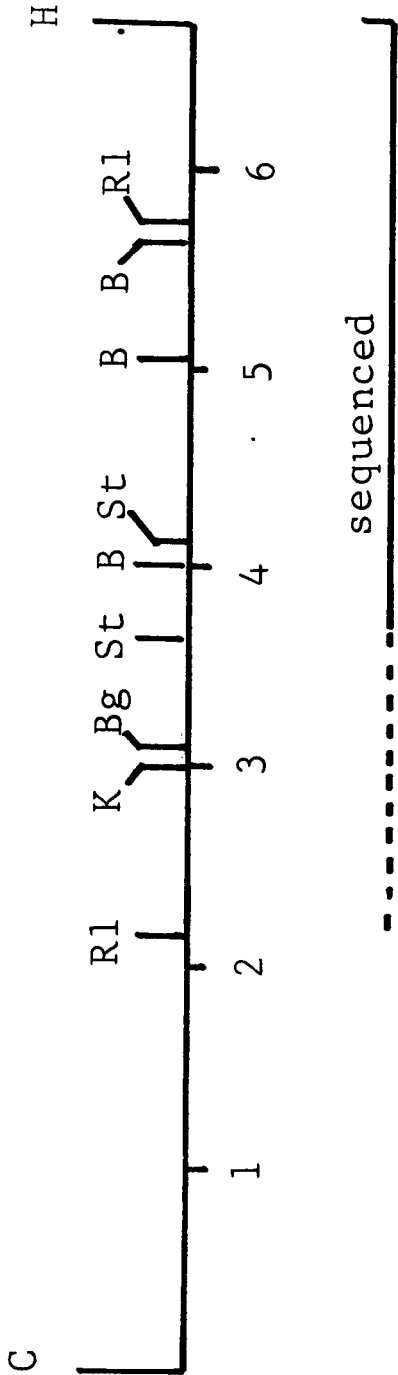
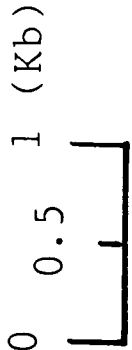
Legend to Fig. 3.3

Restriction map of pCHS2-10

Restriction map was deduced and confirmed as described in Fig.

3.2. Abbreviations: B, BamH I; Bg, Bgl II; H, Hind III; R1, EcoR I; St, Stu I.

Restriction map of CHS2-10

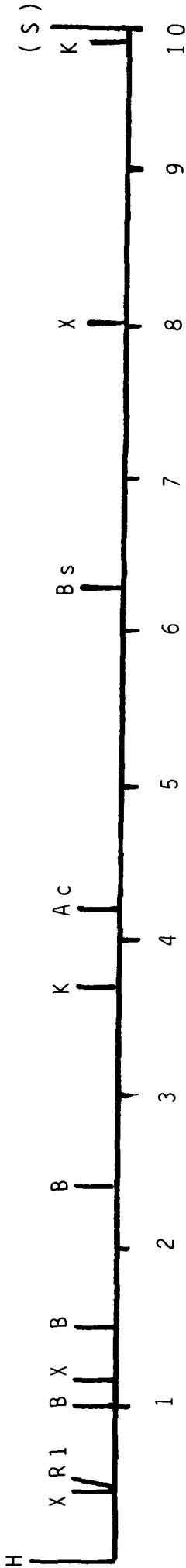


Legend to Fig 3.4

Restriction map of pCHS 1-1

Restriction map was deduced and confirmed as described in Fig 3.2. Two smaller subclones, pAB-3 (containing a 2.1 kb Acc I - BstE II fragment) and pAB-B7 (containing a 3.4 kb BstE II - Sal I fragment) were constructed. Abbreviations: Ac, Acc I; B, BamH I; Bs, BstE II; H, Hind III, K, Kpn I; R1, EcoR I; S, Sal I (vector); X, Xho I.

Restriction map of pCHS1-1



pAB-3

pAB-B7

Legend to Fig. 3.5

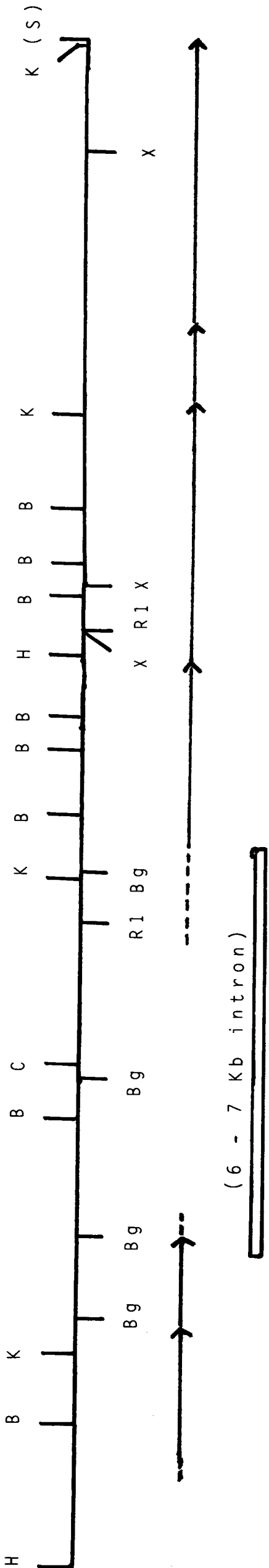
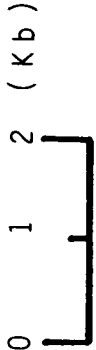
Restriction map of a C4A gene (from Cos 3A3)

Restriction sites were mapped using plasmid-subcloned DNA.

Arrows represent the regions that have been fully sequenced.

Dotted lines represent areas where partial DNA sequences were obtained. The 6-7 kb intron is indicated. Abbreviations: B, BamH I; Bg, Bgl II; C, Cla I; H, Hind III; K, Kpn I; R1, EcoR I; S, Sal I (vector); X, Xho I.

Restriction map of a C4A gene



subcloned DNA, a more complete restriction map of the C4A gene was constructed and shown in Fig. 3.5.

To prepare DNA for M13 sequencing, restriction fragments listed in Table 3.1 were prepared. They were sonicated to generate randomly sized fragments, end-filled to form blunt-end pieces, size fractionated through a 1.5% agarose gel and then ligated to Sma I-cut and phosphatased M13 mp9 or mp8 vectors (see section 2-10).

After DNA sequencing, gel-readings were overlapped and aligned by the computer programmes BATIN, DBAUTO and DBUTIL. The result was the formation of one or more contigs in different databases. Gaps between the contigs in each database were linked by subcloning and sequencing specific DNA fragments.

Altogether 6 DNA contigs were obtained (as shown by the arrows in Fig. 3.5^{Table 3.1}). Besides the gap for the 6-7 kb intron, these contigs were tentatively joined because they represent the ends of the mapped, uncommon restriction sites and they are located within introns. To ensure that there are no small DNA fragments missed due to the presence of identical restriction site(s) in tandem and very close to each other, further subclones and DNA sequence analysis of fragments overlapping the ends between the contigs will be necessary. The complete DNA sequence determined and the protein sequence encoded by each exon is presented in Appendix I.

3.3 Results and Discussion

3.3.1 The 5' regulatory region

The 5' flanking sequence containing the putative C4 promoter is shown in Fig. 3.6. It also includes the first exon that

Table 3.1 DNA fragments generated by restriction digests of various plasmids used to construct the sequence databases

Contig No.	Restriction frgment	Fragment size (kb)	plasmid name	nucleotide No.*
1	SmaI - BglII	2.9	pBH4	A* -600 to 1736
2	BglII - BglII	1.3	pCHS2-1	A* 1737 to 2581
3	Asp718 - HindIII	3.6	pCHS2-10	B* 1 to 3218
4	HindIII - AccI	4.2	pCHS1-1	B 3219 to 7490
5	AccI - PvuII	1.2	pCHS1-1	B 7491 to 8679
6	PvuII - (SalI)	4.4	pCHS1-1	B 8680 to 13048

Plasmid subclones of Cos 3A3 described in Figure 3.1 were digested with the restriction enzymes shown. The DNA fragments were isolated, subcloned into M13 vector by the random method as described in section 2.10.2a and used to generate the nucleotide sequences in the contigs. The nucleotide number* accords to A and B, representing two different databases, shown in Appendix I. Part of the sequence is not included when indicated by *. (Sal) represents the enzyme site present in the cosmid vector.

Legend to Fig. 3.6

The 5' flanking region of the C4A gene from Cos 3A3

Transcriptional start sites are marked by asterisks and the major start site (CAP site) is assigned as nucleotide +1. Three TATA sequences which are located within 9-12 bp repeats are underlined. These sequences are located outside the promoter (i.e. nucleotides -147 to +43; L.C. Wu, unpublished) and their functions are unknown. The enhancer core consensus sequence is overlined. Within the 5' untranslated region are four tetranucleotide repeats, CAGA (double underlined). Amino acid sequence is shown above the DNA sequence. The first exon codes for the 51 bp 5' untranslated sequence, the 19-residue leader peptide with a central hydrophobic core, and the first three residues of the β chain (arrowed). The transcriptional start sites are marked by *.

GCCTGTGATATTTTCTGGATGTCCTTTATTTACTGTGACGTGTGTTTGGGTGCCTTGTTT -561

ATGGGGTAGAGGTGAAGTCTGAGCTTTGCCTCATTCAGAGAGGAAAGGGGTCAGGGGTTT -501

ACTCTGACGTTTCAGGCCATTCTCCCTCGTGGAGTGGTGAGGGTGTACCTAATCTCCTAAA -441

CCACGGAATTTCTGTTAGGGCCTAAAAAAGCAAAAGCCTAGTATAGTTCAATTTGTGTTG -381

GAATGAAAGTAAGAGACAAGTGTCTTAGAAGCCTGTCATTGTTTTGTGAGGGCCTTTAAA -321

TATCCTGTACTCGTGGGCCATGTTGGGCCCTTGACGCCCAGGTATACATGAGCTTGTGT -261

ACCI

GCACCTATACCCTGATACAGATATACCTGGTAGGGGGAGGTGCTCAGGCACTGGAATGAG -201

NcoI

AGGAGTTAACGGGGAAGGACAGGGTTATTTCTGGGCCAAGATTCAGAGTTTCCCATGGAC -141

BanI

MstII

ACCCAGGTGTCCGGGGTGCCCCCACAACCTCTGGGCCTGAGGCCAGTTGCACTTCTTGGCT -81

Enhancer core

DdeI

HincII

GTCACGTGGTTTCCCAGCTTAGCTGGGCTGGGGGAGGAGCAAGGTCCAGAGTCAACTCTG -21

StuI

+1

**

*

CCCCGAGGCCTAGCTTGGCCAGAAGGTAGCAGACAGACAGACGGATCTAACCTCTCTTGG 40

BamHI -19

M R L L W G L I W A S S F F T L

ATCCTCCAGCCATGAGGCTGCTCTGGGGGCTGATCTGGGCATCCAGCTTCTTCACCTTAT 100

→ β

1

S L Q K P R

CTCTGCAGAAGCCCAGGTCCTGGAGGCGGGATGCTGGGTGCTTGGATTGGGGCAGGGCTG 160

encodes the leader peptide and the amino-terminus of the mature C4 protein. This regulatory sequence was characterised in collaboration with L.C. Wu and K.T. Belt. Nuclease S1 mapping and primer extension experiments suggested that there are multiple transcription start sites of C4 messages (which are marked by asterisks in Fig. 3.6). The major start point (i.e. CAP site) was assigned as nucleotide +1. This CAP site is 51 nucleotides 5' to the translation initiation codon. Secondary transcription start sites were found at nucleotides +2 and +19. The putative AT-rich sequence (i.e. TATA box) and CAAT box at position -30 and -70 respectively, which are present in many eukaryotic genes (Breathnach and Chambon, 1981; Dierk et al., 1983), have not been found in complement C4. However, between nucleotides -278 to -230 there are three 9-12 nucleotide repeat regions each of which contains a TATA. The role of these sequences are not known. About 75 nucleotides 5' to the CAP site is a consensus enhancer core (Weiher et al., 1983), which suggests a transcription enhancer may be present.

Through the construction of a series of hybrid genes that contain various C4 5' DNA linked to the bacterial chloramphenicol acetyl transferase gene (CAT) for a sensitive transient expression assay (Gorman et al., 1982), L.C. Wu has shown that the C4 gene contains a relatively strong and compact non-TATA boxed promoter within nucleotides -147 to +43. This promoter region also confers the liver-specific expression of C4 as high transcription activity was observed in a hepatoma cell line but not in an epithelial cell lines after DNA transfection. An enhancer between nucleotides -261 to -13 was

defined, which increased the transcriptional activity of a heterologous SV40 promoter in an orientation and distance independent manner in the CAT assay system.

Within the 5' untranslated region there are three (or probably four) tetranucleotide tandem repeats, CAGA. Whether these sequences play a role in post-transcriptional regulation has yet to be determined.

3.3.2 Exon/intron structure of the C4A gene

The C4A gene of Cos 3A3 spans ~22 kb. The coding sequence of the gene is split by 41 introns into 42 exons (Fig. 3.7). Three of the introns are exceptionally large in size, ie. intron 10 (between exon 10 and exon 11; ~7 kb), intron 33 (1049 bp) and intron 36 (1494 bp). The size of all other introns are more or less comparable to their neighbouring exons. The size of the coding exons varies between 19 bp (exon 6, encodes 6 amino acids) to 232 bp (exon 30, encodes 78 amino acids). The first exon includes 51 bp 5' untranslated sequence, while the last exon include 140 bp 3' untranslated sequence. These 42 exons altogether encode the 1745 amino acid residues of the pre-pro-C4 molecule. The size of exons and introns and the number of amino acid residues encoded by each exon is listed in Table 3.2. The exon phase (refers to section 1.3) and the relevant protein function/structure encoded by each exon are shown in Table 3.3.

3.3.2.1 Structure of the β chain

The first 19 amino acid residues of the precursor molecule form a characteristic leader peptide for a secretory protein. This leader peptide contains a central hydrophobic section with charged amino acids at both ends. Exon 1 also encodes the first three amino acid residues of the beta chain. The beta chain is

Table 3.1 Exon-intron size of a human C4A gene

Exon No.	Amino Acid (aa) encoded		Size (n.t.) ⁺	
	Pro-C4	No of aa residues	Exon	Intron
1	-19 to 3	21	51(5'n.c.)*+65	132
2	4 - 69	66		217
3	70 - 136	67		209
4	137 - 160	24		78
5	161 - 190	30		220
6	191 - 196	6		181
7	196a - 217 (+1)	22		170
8	218 - 250	33		138
9	251 - 285	35		97
10	286 - 329	44		~7 kb
11	330 - 369	39	116	98
12	370 - 428	60	181	145
13	429 - 489	61	182	132
14	490 - 551	62	186	154
15	552 - 604	53	159	252
16	605 - 646	42	127	168
17	647 - 672	25	75	260
18	673 - 737	66	198	90
19	738 - 775	38	113	89
20	776 - 798	23	71	261
21	799 - 845	47	140	113
22	846 - 915	70	210	245
23	916 - 932	17	51	101
24	933 - 962	30	90	135
25	963 - 1032	70	210	179
26	1033 - 1058	26	76	160
27	1059 - 1110	52	157	95
28	1111 - 1149	39	114	107
29	1150 - 1206	57	173	225
30	1207 - 1284	78	232	82
31	1285 - 1340	56	168	395
32	1341 - 1360	20	60	1049
33	1361 - 1292	31	94	97
34	1393 - 1451(+3)	63	186	115
35	1452 - 1481	30	91	90
36	1482 - 1506	25	75	1494
37	1507 - 1540	34	103	164
38	1541 - 1570	30	90	86
39	1571 - 1603	33	99	183
40	1604 - 1631	28	85	263
41	1632 - 1676	45	133	144
42	1677 - 1722	46	143+140 3'n.c.	-

Total No. of aa residues encoded = 1745

* n.c. - non-coding

+ n.t. - nucleotide(s)

Table 3.3 Exon phase and relevant protein structure/function
of the human C4 gene

Exon No.	Exon phase*	Residue(s) of the s.j.	Remarks
1	-2	R	5' nc; leader peptide; N-terminus of the β -chain
2	2 - 0	Q, V	
3	0 - 1	R	
4	1 - 0	E	
5	0 - 2	E	
6	2 - 0	I, I	smallest exon (6 aa)
7	0 - 1	Y	glycosylation site
8	1 - 1	R	
9	1 - 0	K, L	
10	0 - 1	G	
11	1 - 0	Q, A	
12	0 - 0	S, V	
13	0 - 0	M, I	
14	0 - 0	K, L	
15	0 - 0	K, V	
16	0 - 1	K	
17	1 - 1	K	β - α junction ; C4a
18	1 - 1	R	C1s cleavage site ; C4a
19	1 - 2	I	
20	2 - 2	K	
21	2 - 0	T, V	glycosylation site
22	0 - 0	E, K	
23	0 - 1	L	
24	1 - 1	T	2nd factor I cleavage site
25	1 - 1	K	thiolester site
26	1 - 2	W	
27	2 - 0	Q, G	isotypic site
28	0 - 0	V, E	
29	0 - 1	G	
30	1 - 0	Q, D	biggest coding exon (78 aa)
31	0 - 0	Q, F	glycosylation site; 1st factor I cleavage site
32	0 - 0	K, V	
33	0 - 1	T	glycosylation site
34	1 - 2	W	sulphation site; postsecretory proteolytic site; α - γ junction
35	2 - 0	K, L	
36	0 - 0	S, V	
37	0 - 1	P	
38	1 - 1	E	
39	1 - 1	Y	
40	1 - 1	F	
41	1 - 2	H	
42	2 -	-	C-terminus of γ chain; 3' nc

* see Section 1.3 and Patty (1987) for definition

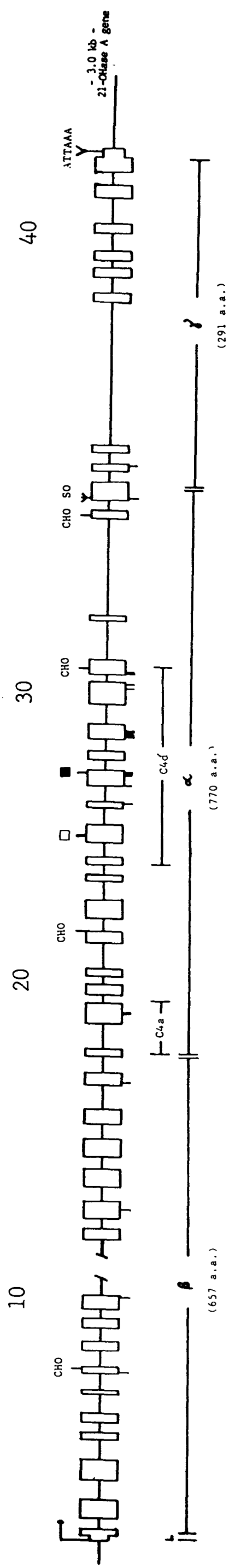
s.j. splice junction; if the s.j. interrupts a codon, the corresponding amino acid encoded by this codon is shown; if the s.j. occurs between two codons, the residues encoded by these two codons are shown.

nc non-coding sequence

Legend to Fig. 3.7

Exon-intron structure of the C4A gene from Cos 3A3

The C4A gene contains 42 exons (open boxes) which code for the 1745 residues precursor C4. The intracellular proteolytic cleavage sites for the $\beta-\alpha$ and the $\alpha-\gamma$ junctions are encoded by exons 17 and 34, respectively. The four glycosylation sites are encoded by exons 7, 21, 31 and 34. The covalent binding site (i.e. the thiolester bond) is encoded by exon 25, and the isotypic residues which may modulate the thiolester reactivity is encoded by exon 27 (see chapter 4 for further description). The three sulphation sites and the post-secretory metalloprotease cleavage site are both encoded by exon 34. $\rightarrow$, transcriptional start site and direction of transcription; CHO, glycosylation site; $\square$, thiolester site; $\blacksquare$, isotypic residues; L, leader peptide; ATTAAA, polyadenylation signal. The polymorphic sites in the coding sequence are represented by vertical strokes underneath the exons.



encoded by exon 1 to exon 17. Genomic sequence suggested that the size of the beta chain is 657 amino acid residues. Compared with the cDNA sequence, an additional Ile residue was found after position 196 and was therefore assigned as number 196a. The missed residue in the cDNA could either be due to a polymorphism in size or due to a cloning artefact.

Compared with cDNA and partial protein sequence data, three polymorphic sites were detected (Fig. 3.8). They are Tyr/Ser 328 (exon 10), Val/Ala 399 (exon 12) and Cys/Ser 616 (exon 16). The Tyr/Ser 328 polymorphism is detected by this work. The cDNA sequences from pAT-A and pAT-F both give Ser 328. Val 399 is present both in Cos 3A3 and pAT-F, while Ala 399 is present in pAT-A. Ser 616 was only detected by protein sequencing (Law and Gagnon, 1985) and has not been found by nucleotide sequencing so far.

The single N-linked glycosylation site on the beta chain, which is of high mannose type (Chan and Atkinson, 1985), is encoded by exon 7 (Fig. 3.9a). The 5 half-cystine residues are encoded by exon 2, 3, 14, 16 and 17 respectively. One of these residues forms the beta-alpha interchain disulphide linkage and the others to form intra-chain bridges (Janatova, 1986; Seya et al., 1986).

The most striking features of the exon/intron structure of the beta chain are the presence of a big intron 6-7 kb in size, and a very small exon with only 19 nucleotides. This 6-7 kb intron has been found in almost all C4 genes located in the first C4 gene locus (generally encodes C4A) as well as in some of the C4 genes located in the second locus (generally encodes C4B) (see chapters 4 and 6). It is of considerable interest

Table 3.4 Comparioson of the polymorphic amino acid residues among the sequences of three cloned C4 genes or cDNAs and the published C4 protein sequences

prot. chain	Exon No.	Position	C4A3a	Allotypes ¹		Prot. seq. ²
				C4A4	C4B2	
β	7*	196a	I	del	del	
	10	328	Y	S	S	
	12	399	V	A	V	
	16	616	C	C	C	S ^a
α	18	707	L	P	P	P ^b
	18	708	D	D	D	N ^b
	26	1054	D	D	D	G ^c #
	27	1090	S	S	S	I ^d
	27	1101	P	P	L	
	27	1102	C	C	S	
	27	1105	L	L	I	
	27	1106	D	D	H	
	29	1157	N	N	S	
	29	1182	T	S	T	
	29	1188	V	V	A	
	29	1191	L	L	R	
	30	1267	A	S	A	
	30	1281	R	R	R	V ^d
	34*	1399a	D	del	D	
	34*	1399b	Y	del	Y	
	34*	1399c	E	del	E	
γ	35	1478	Y	D	Y	

¹ as explained in chapter 4, the sequences of C4A3a, C4A4 and C4B2 are from Cos 3A3, PAT-A and PAT-F, respectively.

² only those polymorphic residues not detected by nucleotide sequencing are shown.

* denotes a probable size polymorphism in the protein

* Gly 1054 has been detected in other allotypes (eg. C4B3 & C4A1) by DNA sequencing; see chapter 4.

del, deleted

^a Law and Gagnon, 1985

^b Moon et al., 1981

^c Chakravarti et al., 1983

^d Chakravarti et al., 1987

Legend to Fig. 3.8

Location of the known polymorphic sites in human C4A/C4B

Three polymorphic residues were detected in the β chain, fourteen in the α chain and only one in the γ chain. There are potential size variations by one residue in the β chain, and by a tripeptide that contains a sulphation site in the α chain.

Polymorphic sites in the α chain are clustered in the C4a (2) and the C4d (12) domains. Polymorphic residues detected by amino acid sequencing alone (not detected by DNA sequencing) are marked by a dot underneath. ∇ , deletion; $\blacksquare$, thiolester site; $\downarrow$, post-secretory metalloprotease cleavage site. (The exact location of the polymorphic residues are listed on Table 3.4).

Fig. 3.9 Exons coding for glycosylation sites

The high mannose-type glycosylation site of the γ chain is encoded by exon 7 (a). The three biantennary complex type glycosylation sites of the α chain are encoded by exon 21 (b), exon 31 (c) and exon 34 (d), respectively. The consensus amino acid sequence for the N-linked glycosylation sites are boxed. The glycosylation site in exon 21 matches the exon-intron splice junction, and that in exon 33 is located just in front of the Cys 1375 which is involved in the formation of a α - γ interchain disulphide linkage. The Breathnach and Chambon rule of intron splicing sequence (i.e. GT-AG) are underlined. The first residue encoded by exon 7 (marked by +) has not been detected by cDNA sequencing. Nucleotide numberings (A): the nucleotide of the CAP site as A1, which applies to the DNA sequence 5' to the 6-7 kb intron; (B) the first nucleotide of the restriction enzyme Stu I recognition site just 3' to the 6-7 kb intron as B1. The same convention applies to Figures 3.11, 3.12, 3.13 and 3.15.

+ 197
 I S A A
 GCATTGCTTCATCAGGATAAATCATATTGTGTCCTTTAGTCAGAAACGCAAGATCTCAGC 1741
 (EXON 7) 217
 R F S D G L E S **N S S** T Q F E V K K Y
 a. CCGATTCTCAGATGGCCTGGAATCCAACAGCAGCACCCAGTTTGAGGTGAAGAAATATGG 1801
 TGAGAGCTGGAAACTGGAGGGACAGGCAGCTGCTTTCCTGAAGGAAATAAGGGTGGAAGG 1861

799 (EXON 21)
 G L C V A T P B
 CTGATCTCTGCCCCTGTGCCCACGCTTCCTTTCCCTCAGGCCTATGTGTGGCCACCCACAG 3300
 b. V Q L R V F R E F H L H L R L P M S V R
 TCCAGCTCCGGGTGTTCCGCGAGTTCCACCTGCACCTCCGCCTGCCCATGTCTGTCCGCC 3360
 R F E Q L E L R P V L Y N Y L D K **N L T** 845
 GCTTTGAGCAGCTGGAGCTGCGGCCTGTCCTCTATAACTACCTGGATAAAAAACCTGACTG 3420

1285
 D B
 GGGGTGGGTAGTCCTCAGACCAAGGGCTTGCTTGAGTCCTGGCTCAACCTCCCTAGGACA 6180
 (EXON 31)
 T V I A L D A L S A Y W I A S H T T E E
 c. CCGTGATTGCCCTGGATGCCCTGTCTGCCTACTGGATTGCCTCCCACACCACTGAGGAGA 6240
 R G L **N V T** L S S T G R N G F K S H A L
 GGGGTCTCAATGTGACTCTCAGCTCCACAGGCCGGAATGGGTTCAAGTCCCACGCGCTGC 6300
 1340
 Q L N N R Q I R G L E E E L Q
 AGCTGAACAACCGCCAGATTGCGGCCTGGAGGAGGAGCTGCAGGTGAACCACTCCCTGG 6360

1361 (EXON 33)
 V L R T B
 TGTGGCCTCACATGTCCCACGTCCTCTCCCTGCCCCTTGCTGAGCCAGGTCCTTCGTACC 7860
 Y N V L D M K **N T T** C Q D L Q I E V T V
 d. TACAATGTCCTGGACATGAAGAACACGACCTGCCAGGACCTACAGATAGAAGTGACAGTC 7920
 1392
 K G H V E Y T
 AAAGGCCACGTCGAGTACACGAGTGAGTGTGGGGGTTGGGAGGCCTTGGGGCCAGGCAGG 7980

Fig. 3.10 The long terminal repeating sequences flanking
 the 6-7 kb intron

a). Comparison of the DNA sequences at the 5' (5-IN.seq) and the 3' (3-IN.seq) ends of the 6-7 kb intron. Dissimilar nucleotide is denoted by -.

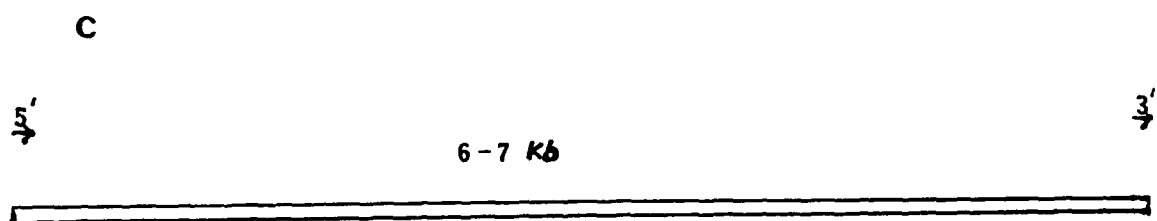
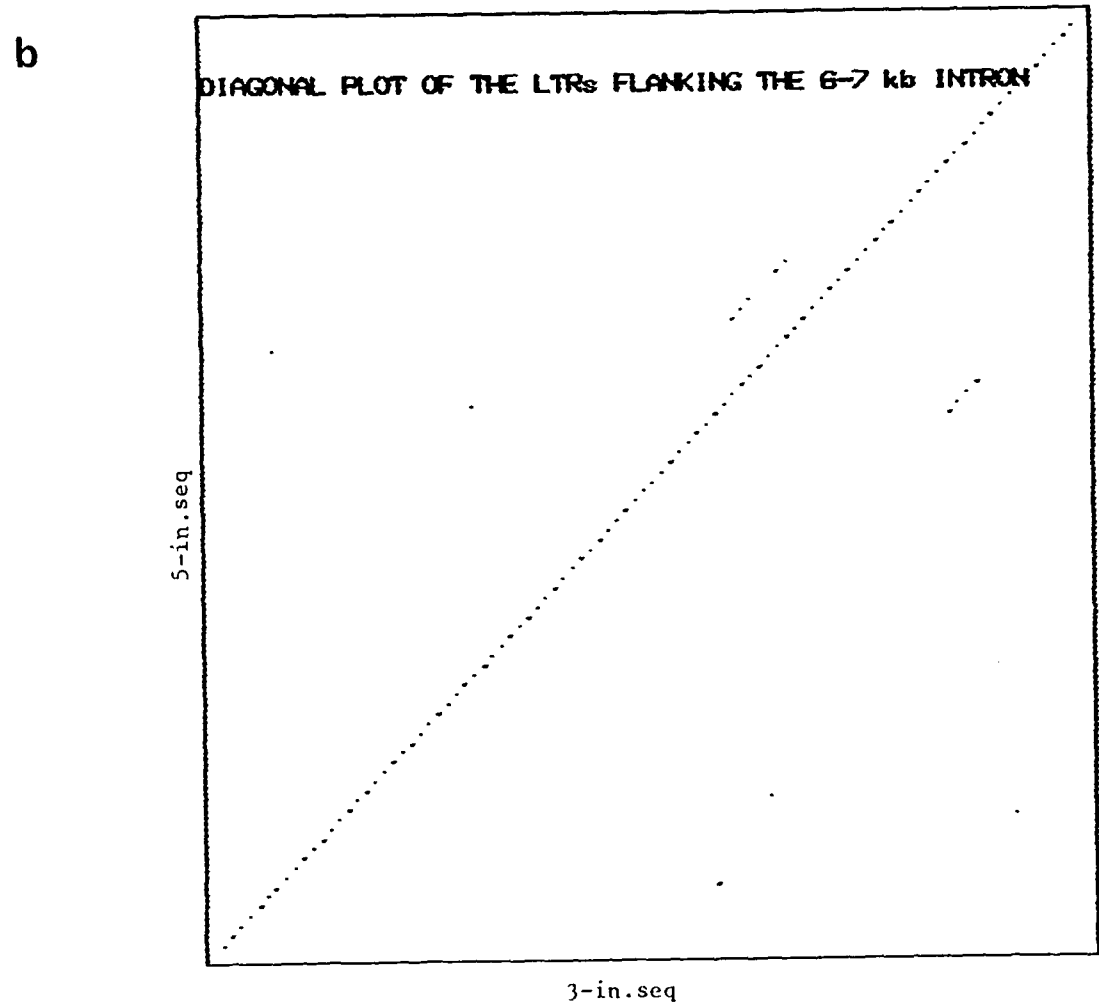
b). Diagonal plot of the sequence shown in a (window size = 13).

c). Location of the direct (denoted by →) LTRs with respect to the 6-7 kb intron (open bar).

a

		10	20	30	40	50
1	5-IN. SEQ	CCTCAAGTCCATTGCTGACACAGCAATGACAACACACAGCTTAACAGCTTC				
2	3-IN. SEQ	CCTCAAGTCTGCTGCTGACAAACCAATGACAACACACAGCTTAACAGCTTC				
	CONSENSUS	CCTCAAGTC—TGCTGACA—ACCAATGACAACACACAGCTTAACAGCTTC				

		60	70	80	90	100
1	5-IN. SEQ	ATAAACAGTGGGGGGCAGGGGGCAATTGCAAAATGCAGCCTGCAAAAGC				
2	3-IN. SEQ	ATAAACAGTGCAGGGCAGGGGGCAATTGCAAAATGCAGCCTGCAAAAGC				
	CONSENSUS	ATAAACAGTGG—GGGGCAGGGGGCAATTGCAAAATGCAGCCTGCAAAAGC				



to determine the DNA sequence of the C4 gene at locus II (or C4B) without this big intron at the corresponding region. Partial DNA sequence revealed long terminal repeats at the 5' and 3' boundaries of this big intron (Fig. 3.10). It is therefore tempting to speculate that this 6-7 kb intron could be a retrovirus-like or a transposable element. This special intron could be involved in mediating the high degree of polymorphism of the C4A / C4B and their neighbouring 21-hydroxylase genes (Yu et al., 1986; Rodrigues et al., 1987).

No known function of the beta chain has been reported. The high mannose glycosylation pattern at position 207 suggests that at least part of the beta chain is buried into the core of the mature C4 protein, rendering further modification of the carbohydrate in the Golgi apparatus impossible (Chan and Atkinson, 1985). Monoclonal antibodies against the C4 beta chain behave like the C3 convertase nephritic factor (Gorski and Muller-Eberhard, 1980; Ichihara et al., 1986). This would infer that the beta chain might be involved in the functioning of the complex enzyme (i.e. C4b2a).

3.3.2.2 Structure of the alpha-chain

The beta-alpha and alpha-gamma proteolytic cleavage sites are encoded by exons 17 and 34 respectively. Most of the known properties / functions of C4 have been ascribed to the alpha chain. The corresponding exons encoding these functions are as follows:

- a. the C1s cleavage site (encoded by exon 18) to generate a biologically active peptide, C4a, which is a weak

Legend to Fig.3.11

Exons coding for the β - α junction and the C4a anaphylatoxin

The proteolytic cleavage site for the β - α junction (boxed) is encoded by exon 17. The first 11 residues of C4a are encoded by exon 17 and the remaining 66 residues are encoded by exon 18. The six Cys residues which may form the three intra-chain disulphide knots are circled. The carboxyl terminus of C4a, i.e. the C1s cleavage site, occurs at the exon-intron splice junction. Both exons 17 and 18 are phase 1-1 symmetrical exons.

647 (EXON 17)

R L S C P K E K T T R K B

GGACTGCAGCCTTTCTCTCTACCAGGACTAAGCTGTCCCAAGGAGAAGACAACCCGGAAA 2160

α

K R $\xrightarrow{C4a}$ N V N F Q K A I N E K 672

AAGAGAAACGTGAACTTCCAAAAGGCGATTAATGAGAAATGTGAGTTGCGGGTGCCTAGG 2220

CAGTAGCTTGGGCTCTCCACCTGGGATCCGGGTGGGGGTCTGCCTCTCTGCCCCCTCGGC 2280

TCCTTGCTGAACCCACGTGTGGTATTTGGGGCCAGAGATCCGAATTCCGGGATTACGAGT 2340

GGAAGGTGGGCAGCTCTCTCCAGCAGCCTCTCTTATGTTGCTGGTCTCAAGGGGTGCGGG 2400

CGGGGGCTGAGGTGTATGTCCTTTTTGTCCTCTCATGCTCACCCCCACCTGGCCCTGCAG 2460

673 (EXON 18)

L G Q Y A S P T A K R C C Q D G V T R L

TGGGTCAGTATGCTTCCCCGACAGCCAAGCGCTGCTGCCAGGATGGGGTGACACGTCTGC 2520

P M M R S C E Q R A A R V Q Q L D C R E

CCATGATGCGTTCCTGCGAGCAGCGGGCAGCCCCGCTGCAGCAGCTGGACTGCCGGGAGC 2580

P F L S C C Q F A E S L R K K S R D K G

CCTTCCTGTCCTGCTGCCAATTTGCTGAGAGTCTGCGCAAGAAGAGCAGGGACAAGGGCC 2640

$\downarrow$
C18
737

Q A G L Q R

AGCGGGCCTCCAACGAGGTAGGGGCTGGGTGGGGCTAGGGCACAGGTGGCGGCGCTTGC 2700

C4a

anaphylatoxin (reviewed by Hugli, 1986). This cleavage also activates C4 and enables the non-covalent interaction between C4b and C2.

- b. two factor I (with C4bp or CR1 as cofactor) cleavage sites (encoded by exons 24 and 31) to generate the C4c and C4d moiety from C4b. This cleavage inactivates C4b and therefore the C3 convertase of the classical pathway.
- c. the thiolester site (encoded by exon 25) in the nascent C4b becomes labile and readily reacts with nearby molecules to form covalent, ester or amide bonds from its reactive carbonyl group.
- d. the C4A / C4B isotypic site (encoded by exon 27) that may modulate the thiolester reactivity (chapter 4).
- e. the Rodgers and the Chido antigenic determinants (encoded by exons 26, 27 and 29, please refer to chapters 4 and 5).
- f. three biantennary complex type, N-linked glycosylation sites (encoded by exons 21, 31 and 33; Fig. 3.9 b,c,d).
- g. three sulphation sites (encoded by exon 34).
- h. a putative metalloprotease cleavage site that may generate a very hydrophilic peptide of 22-26 amino acid residues from the carboxyl terminus of the alpha chain (exon 34).
- i. non-covalent binding sites for C2 (C2a and C2b), CR1, DAF and MCP (location unknown).

Structural and functional correlation of C4a

The weak anaphylatoxin, C4a, is encoded by exon 17 (11 amino acids) and exon 18 (66 amino acids). The C1s cleavage site on native C4 is positioned just 5' to the splice junction of intron

18 (Fig. 3.11). The carboxyl end of this anaphylatoxin contains the active (binding) site of C4a. C4a shares ~30% sequence homology with C3a and C5a and all 6 half-cystine residues are conserved at the same location. Thus these molecules may share similar 3-dimensional structure (reviewed by Hugli, 1984; 1986), and the X-ray crystallographic data from C3a may be a useful model for the 3-dimensional structure of C4a. This would suggest the 3 intra-chain disulphide linkages, i.e. Cys 683 - Cys 709, Cys 684 - Cys 716 and Cys 697 - Cys 717 are formed in native C4a to give a rigid internal core framework, on which the alpha-helical, N terminal and C-terminal regions fold to form a drumstick-like tertiary structure (Huber et al., 1980).

One unexpected result is that two polymorphic sites of human C4a have been found. Genomic sequence defined Leu - Asp 707 -708 (this work), cDNA sequences defined Pro - Asp (Belt et al., 1984; 1985), and protein sequencing defined Pro - Asn (Moon et al., 1981) at the corresponding positions. The Leu - Pro 707 polymorphism is due to C to T transition and can be detected by the restriction enzyme Pvu II. It remains to be determined for whether the polymorphism of C4a has any biological significance.

The factor I cleavage sites (Figs. 3.12 and 1.5)

The metastable C4b is inactivated by two factor I mediated proteolytic cleavages on the alpha chain, with C4bp, or CR1 (or MCP, for the first cleavage only) as a cofactor. The first cleavage site is encoded by the phase 0-0, symmetrical exon 31 (Fig. 3.12). This proteolytic cleavage is sufficient to inactivate C4b to iC4b (Nagassawa et al., 1980). The second cleavage site is encoded by the phase 1-1,

Legend to Fig. 3.12

The exon encoding the factor I-mediated proteolytic cleavage sites

The first and second factor I-mediated cleavage sites on C4 are encoded by exon 31 (symmetrical ; phase 0-0) and exon 24 (symmetrical; phase 1-1), respectively. These two proteolytic cleavages inactivate and split C4b into C4d and C4c.

B

GTACAGGTGGTGGGGGATGTGGACAGCAGGACAGGCTGCCAACTTCCCCCATTTCCCCCAG

2nd I ↓ C4d

933 (EXON 24)

D H R G R T L E I P G N S D P N M I P D

ACCACCGAGGCCGGACCTTGGAATACCTGGCAACTCTGATCCCAATATGATCCCTGATG

4140

4200

962

G D F N S Y V R V T

GGGACTTTAACAGCTACGTCAGGGTTACAGGTGGGAGTGGCCCTTTAGTCCCTTCCCAGTG

GGGGTGGGTAGTCCTCAGACCAAGGGCTTGCTTGAGTCCTGGCTCAACCTCCCTAGGACA

(EXON 31)

T V I A L D A L S A Y W I A S H T T E E

CGGTGATTGCCCTGGATGCCCTGTCTGCCTACTGGATTGCCTCCCACACCACTGAGGAGA

1285

D

B

6180

6240

C4d ↓ 1st I

R G L N V T L S S T G R N G F K S H A L

GGGGTCTCAATGTGACTCTCAGCTCCACAGGCCGGAATGGGTTCAAGTCCCACGCGCTGC

1340

Q L N N R Q I R G L E E E L Q

AGCTGAACAACCGCCAGATTGCGGGCCTGGAGGAGGAGCTGCAGGTGAACCACTCCCTGG

6300

6360

symmetrical exon 24 (Fig. 3.12). The second cleavage splits iC4b to C4d and C4c. The 45 000 mol.wt. C4d may bind covalently to nearby surfaces. The C4d fragments constitute to the formation of the Rodgers (Rg) and the Chido (Ch) antigens in the plasma or on the erythrocytes).

The covalent binding site (Fig. 3.13)

The covalent binding, thiolester site in C4 is encoded by the second largest exon of the gene, i.e. exon 25 (70 amino acids). The protein sequence is ~32 -35% homologous to the analogous regions of human C3 and alpha-2 macroglobulin. The thiolester site on C4 corresponds to the middle region of exon 25 (Fig. 3.13) and is surrounded by a hydrophobic pocket. It will be of considerable interest to determine whether those of C3 and alpha-2 macroglobulin are coded by a similar exon, and also to determine the sequence and exon structure of the other evolutionary related complement component C5, which does not , however, possess a thiolester bond. Knowledge to this end may give insights on the origin of this unique, covalent binding region. The covalent binding reactivity of the thiolester carbonyl exhibits differential binding affinities to amino and to hydroxyl groups of the nearby 'substrate' molecules. This differential binding affinity seems to be modulated by the C4A / C4B isotypic region that is ~107 residues C-terminal to the thiolester site. Molecular cloning and sequence determination of relevant, C4 rare variants have enabled the elucidation of the isotypic residues, i.e. PCPVLD 1101-6 for C4A and LSPVIH for C4B at the same position. These isotypic residues are encoded by exon 27 (refer to chapters 4 and 5 for further

Legend to Fig 3.13

Exon encoding the thiolester site

The thiolester site (overlined) is encoded by symmetrical,
phase 1-1 exon 25.

963 (EXON 25)
A S D P L
B

TCTGACACAGAGTCCTCAGACCTCCACCCTCTCCCTCCCATGTAGCCTCAGATCCATTGG4380

D T L G S E G A L S P G G V A S L L R L
ACACTTTAGGCTCTGAGGGGGCCTTGTCACCAGGAGGCGTGGCCTCCCTCTTGAGGCTTC4440

Thioester

P R G C G E Q T M I Y L A P T L A A S R
CTCGAGGCTGTGGGGAGCAAACCATGATCTACTTGGCTCCGACACTGGCTGCTTCCCGCT4500

Y L D K T E Q W S T L P P E T K D H A V
ACCTGGACAAGACAGAGCAGTGGAGCACACTGCCTCCCGAGACCAAGGACCACGCCGTGG4560

1032

D L I Q K
ATCTGATCCAGAAAGGTTCTGGGTGCAAGGGCAAGCAGGAGGGGGGCCAGGAAAGGACAG4620

discussion).

The non-covalent binding site(s) (Fig. 3.14)

C2, C4bp, CR1, decay accelerating factor (DAF) and membrane co-factor protein (MCP) may share related, or overlapping, if not the same, non-covalent binding sites on C4. These sites are ill-defined and have not been properly mapped. However, analogous binding site(s) for factor H and CR2 has been located to a region within residues 1199-1274 on the human C3 molecule (reviewed by Lambris and Muller-Eberhard, 1986). This would suggest that an analogous non-covalent binding site on C4 might be localised on a region mainly encoded by exon 30, if C3 and C4 possess the analogous non-covalent binding site at the similar position (Fig. 3.13). Exon 30 is the largest coding exon of the C4 gene and it codes for 78 amino acids. Its derived protein sequence shares 38% identity over a stretch of 42 amino acids when compared with C3 in this region. No sequence similarity with C3 was found at the N-terminal region of the peptide encoded by this exon, which suggests that the specificity of interaction among C4 and C3 with their analogous but different ligands may lie in this region. The protein sequence encoded by exon 30 is hydrophobic in the middle and flanked by hydrophilic regions at both ends.

A unique region encoded by exon 34 (Fig. 3.15)

C4 is different from C3 and C5 in that the mature molecule is post-transcriptionally processed to a 3-chain structure before secretion, instead of the 2-chain form for C3 and C5. This is because C4 undergoes an extra intracellular proteolytic cleavage. C4 is unusual as it undergoes an

Legend to Fig. 3.14

Analogous non-covalent binding site on C4 (?)

The factor H binding site (dashed underlined) and the CR2 binding site (solid underlined) on the C3 molecule have been mapped (reviewed in Lambris and Muller-Eberhard, 1986). Assuming the analogous non-covalent binding site on C4 (i. e. for C4bp etc.) is located at the equivalent position as C3, this site would map to a region mainly encoded by exon 30. Alignment of the sequences revealed that C4 and C3 share 38% homology in an overlap of 42 residues. (It is possible that there are more than one non-covalent binding sites on the C4 molecule).

∴, identity in sequences; ., conservative changes; -, gap included to maximize homology.

(Exon 30)

DNL YWGS

C3
1141" MALTAFVLISLQEAKDICEEQVNSLPGSITKAGDFLEANYMNLQRSYTVAIAGYALAQMG

C4 VTGQSNAVSTPAPRNPSPMPQAPALWIETTAYALLHLLHECKAEMADQAAAWLTRQ

C3 1201" RLKGPLLNKFLTTAKDKNRWEDPGKQLYNVEATSYALLA-LLQLKDFDFVPPVVRWLNEQ
----- (CR 2) -----

C74
GSFQGGERSQ

C3 1260" RYGGYGSTQATFMVFQALAQYQKDA PDHQELNDVSLQLPSRSSKITHRIHWESALL

Legend to Fig. 3.15

Exon 34 - a unique exon found in C4

Exon 34 encodes three O-tyrosine sulphation sites (circled), the post-secretory metalloprotease site (arrowed) and the intracellular proteolytic cleavage site (boxed) for the α - γ junction. All these features are unique to C4 among the complement proteins. The size-polymorphic tripeptide that contains one of the sulphation site is marked by +. Cp, carboxyl terminus of plasma C4; Cs, carboxyl terminus of secretory C4. (Exon phase = 1-2; non-symmetrical).

1393

M

B

GGCTGGCGCAGGGAGCCGGGTGGCCATCCCAGCCCTCCTCACAATGCTTCCCTGTGCAGT

8040

(EXON 34)

+ + +

E A N E D (Y) E D (Y) E (Y) D E L P A K D D P

GGAAGCAAACGAGGACTATGAGGACTATGAGTACGATGAGCTTCCAGCCAAGGATGACCC

8100

D A P L Q P V T P L Q L F E G R R N

AGATGCCCCTCTGCAGCCCGTGACACCCCTGCAGCTGTTTGAGGGTCGGAGGAACCGCCG

8160

Cs

→

R R

E A P K V V E E Q E S R V H Y T V C

CAGGAGGGAGGCGCCCAAGGTGGTGGAGGAGCAGGAGTCCAGGGTGCACTACACCGTGTG

8220

1451

I W

CATCTGGTGGGCGCCGGGAGCTGCCCTGGGCCAGGGGAGGGAGGGCAGGACCCAGGCTGG

8280

GGCTGGGCTTCTGGAGCCCGCGCAGGCAGAACCTGGACGACAGCTCACACGTCTCCACAG

8340

extracellular proteolytic cleavage by a metalloprotease after secretion (Chan et al., 1984). This additional processing would generate a hypothetically, very hydrophilic peptide 22-26 residues from the C-terminus of the alpha chain (Law and Gagnon, 1985). Moreover, C4 is unique on comparison with its related proteins as the three tyrosine residues near the C-terminal region of the alpha chain are sulphated (Karp, 1983; Hortin et al., 1986a). The intracellular and extracellular proteolytic cleavage sites together with the three sulphation sites are all encoded by exon 34. This exon-encoded peptide has an unusual amino acid sequence. In a region of 12 residues long, there is a proline residue in every tripeptide. About 40% of the 63 amino acids encoded by this exon are charged. Together with the sulphate groups linked to the three tyrosines, these form a highly negative region on the carboxyl terminus of the alpha chain, and a hydrophilic N-terminus of the gamma chain. It was suggested that the C-terminus of the alpha-chain may be involved in holding the Mg^{2+} which is required for the C4b2a (C3 convertase) activity (Hortin et al., 1986a). Indirect evidence has suggested that the C-terminus of the alpha chain is exposed to the exterior as the sulphated peptide can be proteolytically excised and eluted without denaturing the protein (Chan et al., 1983). No sequence homology at this region has been observed between C3, C4 and alpha-2 macroglobulin, suggesting that this exon may be unique to C4.

One further point of interest for this exon is that a tripeptide, DYE 1199a, b, c, (Fig. 3.15) which contains one of the three sulphation sites, is not present in a cDNA clone

PAT-A. This may represent another degree of polymorphism on the C4 molecule.

Cysteine residues and disulphide linkages

There are 11 cysteine residues on the alpha chain encoded by Cos 3A3. Six of them are present on C4a to form a rigid, disulphide 'knot' as described. One cysteine is involved in the formation of the thiolester bond (i.e. Cys 994). Cys 1102 is one of the C4 isotypic residues and may be involved in modulating the thiolester activity. This residue may form some thiolester-like linkage with Asp 1106, or be buried, or substituted, as no free thiols have been detected except the one released from the thiolester, after activation of C4. An alpha-beta and an alpha-gamma inter-chain disulphide linkage was detected on the 27 K, alpha 3 fragment of C4b (Janatova, 1986; Seya et al., 1986b). This suggests that Cys 801 encoded by exon 21 and Cys 857 encoded by exon 22 are involved in linking the alpha chain to the beta and gamma chains. Another interchain disulphide linkage was detected between the 16K alpha 4 fragment and the gamma chain. This implies that the Cys 1375 encoded by exon 33 is involved in the formation of the second alpha-gamma bridge (Fig. 1.5).

Glycosylation sites (Fig. 3.9 b, c and d)

Three N-linked glycosylation sites are present on the alpha chain and they are of biantennary complex type (Chan and Atkinson, 1985). They are encoded by exon 21 (position 843 - 845), exon 31 (position 1309 - 1311) and exon 33 (position 1372 - 1374). The first site matches the splicing junction of intron

21. The last site is located just N-terminal to the cysteine residue which is involved to the alpha-gamma inter-chain disulphide linkage.

Polymorphism

Most of the polymorphic sites on C4 are detected on the alpha chain and they are all clustered at the C4d region except the two present on the anaphylatoxin, C4a, described before. A total of 12 amino acid residues have been detected by protein sequencing of C4 from pooled serum. Genomic sequence of the C4A gene, together with the published cDNA data, in this study, have detected ten of the described amino acid changes (Table 3.4; Fig. 3.8). These changes account for the C4A/C4B isotypic properties, Rg/Ch antigenic determinants and the allelic variation.

3.3.2.3 Structure of the gamma chain

Compared with the those of alpha and beta chains, the exons coding for the gamma chain appear to be more uniform in size. Except for exon 34 (63 residues) which encodes the N terminus of the gamma chain, the other eight exons encode 25 to 45 amino acid residues. Exons 36 and 37 are interrupted by the second biggest intron of the gene, i.e. 1494 bp. No protein function for the C4 gamma chain is known. In terms of secondary structure, the gamma chain may probably be more rigid as it possesses 5 intra-chain and two inter-chain (alpha-gamma) disulphide linkages. The 12 half-cystine residues are encoded by exons 34 (1), 37 (1), 38 (4), 39 (2), 41 (1) and 42 (3). No glycosylation site has been found on the gamma chain. A single polymorphic site has been

detected on the gamma chain, i.e. Asp/Tyr 1478. Tyr 1478 is present in the C4A gene sequenced.

3.3.3 The 3' ends of the C4 genes

The last exon of the C4A gene also includes a 3' untranslated sequence of 140 nucleotides. The polyadenylation signal is a variant ATTAAA instead of AATAAA (Proudfoot and Brownlee, 1978) which is 105 nucleotides downstream from the stop codon. In other words, the poly(A) site is 31 bp 3' to the polyadenylation signal in the C4A gene sequenced. Whereas in the C4B cDNA pAT-F, this distance is only 14 bp (Belt et al., 1985). Immediately downstream of the poly(A) site is a stretch of GT-rich sequence which has been suggested to be involved in the formation of the 3' end of the mRNA (Gil and Proudfoot, 1987).

The DNA sequence between the poly(A) site of the C4 gene and its immediate downstream Bgl II site has also been determined, it is 1357 bp long. Recently, Higashi et al., (1986) published the DNA sequences of the neighbouring 21-hydroxylase A and B genes starting from the Bgl II sites. The DNA sequence of a 0.6 kb Asp 718 restriction fragment corresponding to the 3' end of the C4B 3 gene has been determined (data not shown). This sequence showed that the 3' ends of the C4A and C4B genes are identical. As the mentioned Bgl II site is about 1671 bp 5' to the translation initiation codon of the 21-hydroxylase gene, this suggests that the intergenic region between C4A and 21-hydroxylase A, and presumably between C4B and 21-hydroxylase B, is about 3028 nucleotides. This is considerably longer than the distance separating the other two HLA-linked complement

genes, C2 and Factor B, which is only 421 nucleotides apart (Wu et al., 1987).

3.3.4 Exon phase and amino acid residues encoded at/between the splicing junctions

As shown in Table 3.3, there are 17 symmetrical and 25 asymmetrical exons in the C4A gene. Among the symmetrical exons, nine belong to the 0-0 phase, eight belong to the 1-1 phase and only one belongs to the 2-2 phase. Many of the important functions in the C4 protein are encoded by the phase 1-1 exons. These include the proteolytic cleavage site for the beta-alpha junction, the two exons encoding the anaphylatoxin C4a, the thiolester exon and the exon encoding for the second factor I cleavage site. It has been noted that most of the exons involved in exon shuffling / duplication events in the plasma proteins of vertebrates belong to the 1-1 phase (Patthy 1987). It is notable that an anaphylatoxin is not encoded by the C4 related protein, α_2 macroglobulin; also a thiolester bond is absent in the other C4 related protein, C5. Thus it is tempting to speculate that the absences of the corresponding protein functions in α_2 M or in C5 may be caused by exon deletion or insertion events. The symmetry in the exons favours the exonic recombination events because they would not disrupt the reading frame of the coding sequence which otherwise would be deleterious .

It has been noted by Craik and coworkers (1982, 1983) that the exon-intron splice junctions often map to protein surfaces. In other words, the amino acid residues encoded at or between the splice junctions tend to be hydrophilic. This phenomenon may well apply to the case C4. Remarkably, among the residues encoded by the 3' ends of the 42 exons in C4, 18

are charged and 14 are polar in nature. Of the charged residues, seven are lysine. It is of considerable interest to see whether a similar feature is also present in the other related proteins.

3.3.5 Comparison of the exon/intron structure among human C4, mouse C4 and mouse C3 genes at the 5' region

Limited genomic sequences of murine C4 (Nonaka et al., 1985) and murine C3 (Lundwall et al., 1984) are available. A comparison of their exon/intron structures with that of human C4 is shown in Table 3.5. Human C4 and mouse C4 share exactly identical exon size and exon phase for the regions compared. This is in accord with the phenomenon that there is a temporal conservation of the exon/intron structure (i.e. exon size and intron position) of the same (analogous) protein among different species. On the other hand, the sequence and size of the introns between human C4 and mouse C4 do not shown homology, probably because there is no selection pressure to maintain the identity. Computer analysis using the programme DIAGON has already revealed that there is very little or no homology between C4 and C3 in the N-terminal region of the precursor proteins (Fig. 1.8a). This disparity is again exhibited in the exon/intron structure. The size and phase of the first two exons of mouse C3 are both different from those of C4, although the first exon of mouse C3, mouse C4 and human C4 all encode the leader peptides and the first three residues of the N-terminus of the corresponding precursor protein.

3.4 Summary and conclusion

The genomic sequence of a C4A gene and its 5' and 3'

Table 3.5 A comparison of the exon/intron structure at the 5' end regions among the human C4, mouse C4 and mouse C3 genes

	<u>human C4</u>	<u>mouse C4</u>	<u>mouse C3</u>
exon 1* (nt)	51 nc + 65	51 nc + 65	X + 80
(aa)	22 (19 + 3)	22 (19 + 3)	27 (24 + 3)
phase	- 1	- 1	- 2
exon 2 (nt)	199	199	190
(aa)	66	66	63
phase	1 - 0	1 - 0	2 - 0
intron 1 (nt)	132	100	1061
intron 2 (nt)	217	457	361

nt, nucleotides;

nc, non-coding sequence;

aa, amino acids;

X, 5' non-coding sequence of mouse not determined;

*, exon 1 encodes the 5' nc, the leader peptide and the protein N-terminus.

flanking region has been determined. This allows the elucidation of the exon - intron structure of the C4 gene. C4 has a compact, non-TATA boxed and cell-specific promoter which can be potentiated by a transcription enhancer element whose core is located ~70 nucleotides upstream of the major CAP site. Between the CAP site and the translation initiation codon is a 5' untranslated region of 51 nucleotides.

The precursor C4 molecule is encoded by 42 exons. The beta-alpha and alpha-gamma proteolytic cleavage sites are respectively encoded by exon 17 and exon 34. Genomic sequence suggests that the C4A gene codes for a 1745 amino acids molecule, of which the first 19 residues forms a leader peptide with a hydrophobic core that is characteristic of secretory proteins.

There is some correlation between the exon structure and the protein function of C4. Most of the known protein functions and properties have been located to the alpha chain. The 77 residue anaphylatoxin, C4a, is encoded by the second half of exon 17 and the entire exon 18. The C1s cleavage site is at the last amino acid residue encoded by exon 18, which matches completely with the exon-intron boundary. The thiolester covalent binding site is encoded by exon 25 while the C4A/C4B isotypic residues which may modulate the thiolester reactivity is found in exon 27. The Rodgers and Chido antigenic determinants are encoded by exons 26, 27 and 29. It is probable that the peptide fragment encoded by exon 30 may be involved in the formation of the non-covalent binding site of C4bp and/or other related proteins. Exon 34 appears to be unique to C4 in that it codes for (i) the intracellular proteolytic cleavage site for the alpha-gamma

chain, (ii) extracellular metalloproteolytic site to remove 22 to 26 residues from the C-terminus of the alpha chain, and (iii) it encodes three O-tyrosine linked sulphation sites. These properties have not been found in other related proteins, C3, C5 and alpha-2 macroglobulin. The protein sequence encoded by this exon does not exhibit sequence identity with those of the related proteins at the analogous positions. Obviously, more information about the evolutionary relationship between C4, C3, C5 and alpha-2 macroglobulin will be deduced when the gene organisation of the last three proteins are determined.

Polymorphisms have been detected on all three chains of the C4 molecule (Fig. 3.10). There is a possible variation in size by one amino acid residue for the beta chain (exon 7) and by three residues, one of which also contains a sulphation site, on the alpha chain (exon 34). Genomic, cDNA and protein sequences together have detected three changes in amino acid sequence in the beta chain, fourteen changes in the alpha chain, and one change in the gamma chain. For the alpha chain, two changes were located in the C4a activation peptide and all others are found in the C4d region. These changes seem to be related to the variation in the differential thiolester reactivities antigenicity and electrophoretic mobilities. Polymorphism constitutes one of the most fascinating aspects on the studies of C4.

The size of the 41 introns is generally comparable to their neighbouring exons except for 3 that are more than 1 kb larger. The intron between exon 10 and 11 is about 6-7 kb, which is unusual in that it is absent in some of the C4 genes. Preliminary

data reveal the presence of long terminal repeats flanking the 5' and 3' ends of this intron. It remains to be determined whether this intron is a transposable element .

The last exon of C4 also contains 140 bp of 3' untranslated sequence. The poly(A) signal is found 31 bp upstream the poly(A) site. Near the poly(A) site is a stretch of GT-rich sequence which may be involved in mRNA formation. Together with published sequences for the 21-hydroxylase A and B genes, the intergenic region between C4 and 21-hydroxylase is found to be ~3028 bp.

CHAPTER FOUR

STRUCTURAL BASIS OF THE POLYMORPHISM OF HUMAN COMPLEMENT

COMPONENTS C4A AND C4B: GENE SIZE, REACTIVITY AND

ANTIGENICITY

4.1 Introduction

Human complement component C4 plays a key role in the activation of the classical complement pathway, especially in the formation of the complex C3 and C5 convertases (for reviews see Reid and Porter, 1981; Porter, 1984). There are two C4 isotypes, C4A and C4B, which are encoded by tandem loci about 10 kb apart, located in the major histocompatibility complex on the short arm of chromosome 6 (Teisberg et al., 1976; Carroll et al., 1984a; reviewed by Porter, 1985; Carroll et al., 1985a). As shown by the derived amino acid sequences from complete and/or partial cDNA and genomic sequences, the two protein molecules are very similar and there is less than 1% sequence variation between the pro-C4 molecules with 1745 amino acid residues (Belt et al., 1984; 1985, chapter 3). C4 is synthesised as a single chain precursor molecule (mol. wt . 200,000) and is post-translationally processed to the three-chain disulphide-linked structure (Hall and Colten, 1977; Janatova, 1986).

At the middle of the α chain, there is a characteristic thiolester bond which is also present in complement component C3 and the protease inhibitor, α_2 macroglobulin (reviewed by Tack, 1985). After activation of C4, the thiolester bond becomes labile and is readily cleaved to give a reactive

carbonyl that will form a covalent ester or amide bond with nearby antigens or Ig G surfaces (Law et al., 1980; Campbell et al., 1980). Despite the high sequence homology, C4A and C4B exhibit a remarkable difference in the thiolester reactivity. Recent analyses by two laboratories (Ise^hman and Young, 1984; 1986; Law et al., 1984; Dodds et al., 1986) have independently led to the similar conclusion that C4A binds preferentially to amino groups of peptide antigens and C4B more efficiently binds to hydroxyl groups of carbohydrate antigens. This intrinsic difference in covalent binding affinity results in a 3-4 fold higher binding of C4B to sensitised sheep red cells and explains the well-known higher activity of C4B in the classical haemolytic assay (Adweh and Alper, 1980). The α chain of C4A has a lower electrophoretic mobility of about 2,000 daltons than C4B (i.e. mol. wt.: C4A-96,000 vs. C4B-94,000; Roos et al., 1982). The human blood group antigens Rodgers (Rg) and Chido (Ch) were shown to correspond to C4A and C4B, respectively (O'Neill et al., 1978a). The apparent 2,000 dalton mol. wt. difference of the α chain and the Rg/Ch antigenic determinants reside in the proteolytic degradation fragment, C4d (Tilley et al., 1978; Lundwall et al., 1982; Chan and Atkinson, 1985).

C4A and C4B are very complex both at the genomic and at the protein levels. This may have important implications for the origin of many HLA-related autoimmune diseases (Porter, 1983; Dawkins et al., 1983; Fielder et al., 1983). The C4 genes of different individuals can differ both in number and in size (Carroll et al., 1985b; reviewed by Carroll et al., 1985a; chapter 6). The C4A and C4B proteins are highly

polymorphic and this has been demonstrated by differences in electrophoretic mobility of neuraminidase-treated plasma (Mauff et al., 1983a; Sim and Cross, 1986), serological typing (Roos et al., 1984; Giles et al., 1984; Giles, 1987), and direct DNA sequencing (Belt et al., 1984, 1985; and this work). A combination of all these approaches suggests that there are altogether more than forty alleles in the two C4 classes.

By means of many polyspecific anti-Rg and anti-Ch antisera obtained from transfused patients who lacked either C4A or C4B protein, respectively, it has been defined that there are two Rodgers (Rg: 1,2) and six Chido (Ch: 1, 2, 3, 4, 5 & 6) antigenic determinants (Giles et al., 1984; Giles, 1987). However, there is not a complete association of Rg determinants with C4A and Ch determinants with C4B. Despite maintaining the class-specific differential thiolester reactivities and having the class-specific apparent α chain mol. wt. (Rittner et al., 1984; Roos et al., 1984; Dodds et al., 1986), some C4A 1 allotypes do not possess either of the two Rg determinants, but express four Ch determinants (i.e. Rg:-1,-2; Ch: 1,-2,3,-4,5,6). Moreover, a minor proportion of the C4B 5 allotypes have been shown to express Rg1 (whether Rg2 is express or not has not been determined), and contain only two of the six Ch determinants (i.e. Ch: -1,-2,-3,4,-5,6) (Rittner et al., 1984; Roos et al., 1984; Giles et al., 1984; Giles, 1987). It was not clear whether the Rg/Ch determinants and the differential thiolester reactivities are correlated.

In order to correlate the structure and function of the

TABLE 4.1 Haplotypes of individuals from whom λ EMBL-3 genomic libraries were constructed

Individuals	HLA	C4	Deduced Rg/Ch types	References
1. A.W.	A2 Cw3 B15 DR1 A2 - B7 DR2	A3 B3 A3 BQ0	Rg: 1,2; Ch: 1,2,3,4,5,6 Rg: 1,2; Ch: -1,-2,-3,-4,-5,-6	Carroll et al., 1985a; this work.
2. J.M.	A3 B35 DR4(?) A24 B18 DR2(?)	A3 BQ0 A4 B5	Rg: 1,2; Ch: -1,-2,-3,-4,-5,-6 Rg: 1,2; Ch: -1,-2,-3,4,-5,6	Giles et al., 1984; Hing et al., 1986; Giles, 1987.
3. A.D.	A3 C6 B47 DR7 A2 C5 B44 DR6	A1 BQ0 ⁺ A3a BQ0	Rg: -1,-2; Ch: 1,-2,3,-4,5,6 Rg: 1,2; Ch: -1,-2,-3,-4,-5,-6	Bentley et al., 1985.

+ This was previously typed as C4A Q0; C4B 3.1 but further studies show that it is C4A 1; C4B Q0.

two C4 isotypes particularly with respect to the structural basis of the differential thiolester reactivity, the antigenicity of the Rg/Ch blood groups, and the polymorphism of C4A and C4B, genomic libraries using DNA isolated from three individuals with interesting Rg/Ch phenotypes were constructed^{Table 4.1}. From these libraries genomic clones of the complete Rg+ C4A 3a (C4A 3a is a sub-type of C4A 3) and the complete Ch+ C4B 3 were isolated. In addition I obtained clones for the C4A 1 and C4B 5 allotypes, described above, which possess their own class-specific properties, but essentially the reversed antigenicity. The C4 genes of these clones were characterised and the polymorphic C4d regions sequenced. This work allows many important conclusions to be drawn on the genetics and polymorphism of the C4A and C4B isotypes and advances our understanding of the structure and function of two very similar but unique proteins which probably originated from the same ancestral gene by gene duplication.

4.2 Results

4.2.1

Cloning, mapping and assignment of C4 genes

Hybridisation of full-length and 5' C4 cDNA probes to restriction digests of cloned C4 genes has shown that the C4 genes vary in size. The mapping of a C4A gene in the cosmid Cos 3A3 revealed that the gene is 22 kb in length and has a 6-7 kb intron located near the 5' end of the gene (Carroll et al., 1985a; Carroll et al., 1986). In contrast two different C4B 1 alleles, C4B 1a from Cos 1E2 and C4B 1b from Cos KEM-1 (Carroll et al., 1985a; 1985b; Belt et al., 1985) do not possess the big intron and are only 16 kb in size. A combination of

Southern blot analysis on subcloned DNA using a full length cDNA probe, and direct DNA sequencing of the C4A gene from Cos 3A3 has enabled the construction of an accurate restriction map of the gene (Fig. 4.1a). The 6-7 kb intron is located about 2-2.5 kb from the 5' of the gene. Its presence can be detected by Southern blot analysis of BamH I or Kpn I restriction genomic fragment, using a 476 bp 5' cDNA (P_A) and a 927 bp BamH I genomic fragment specific for the C4d region of the gene (P_B), respectively. The long C4 gene can be identified by the presence of the 4.8 kb Bam HI and the 7.5 kb Kpn I fragments. A short C4 gene is characterised by the presence of the 3.3 kb BamH I and the 8.5 kb Kpn I fragments (Carroll et al., 1985a; 1985b). The genomic DNA of the three individuals whose DNA was used to construct genomic libraries was examined by this approach.

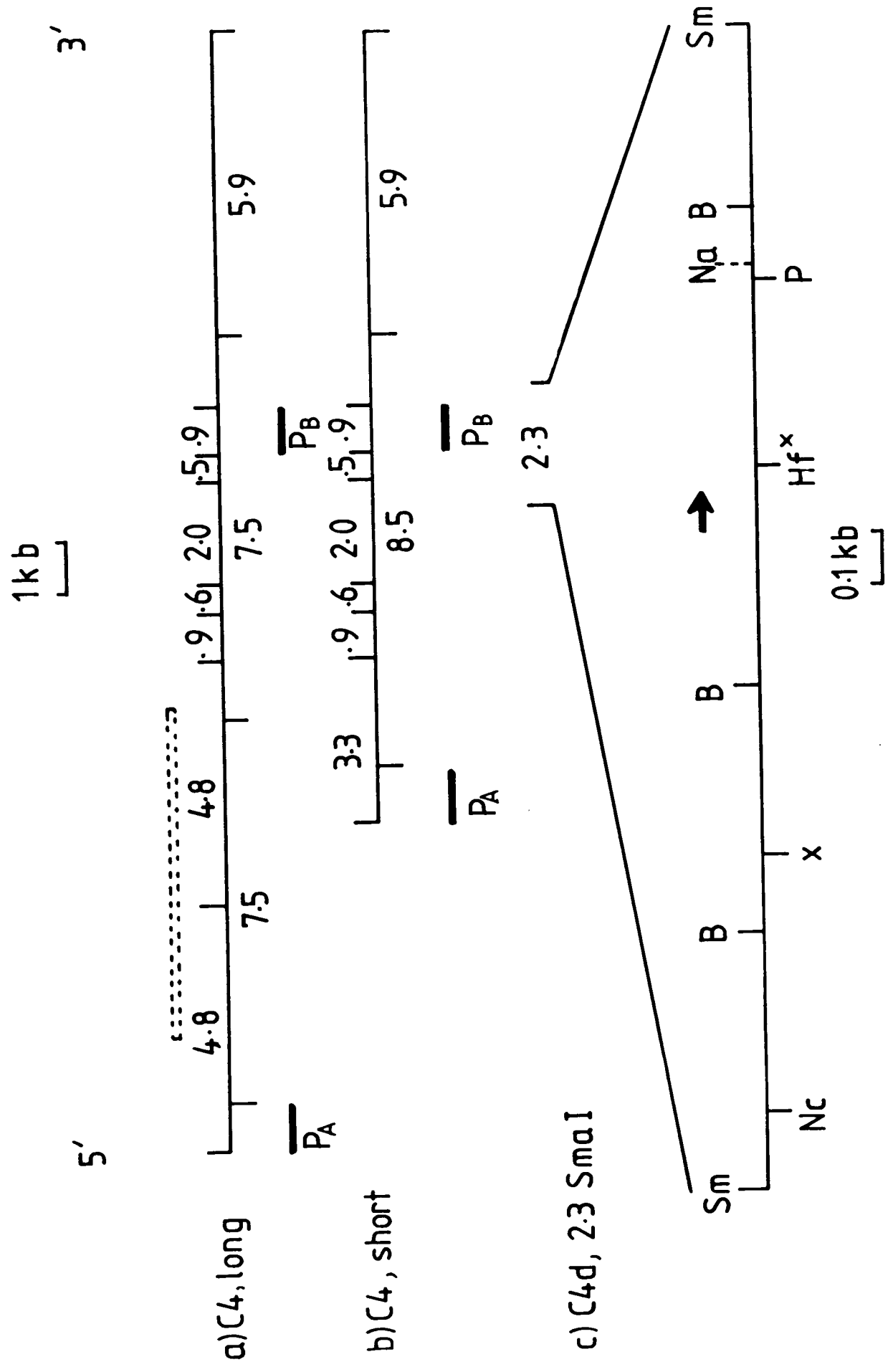
In the individual AW (C4A 3; C4B 3/ C4A 3; C4B Q0), only the 4.8 kb BamH I and 7.5 kb Kpn I fragments were detected (Fig. 4.2, lanes 1, 4). As the C4B Q0 allele in this individual was deleted (Carroll et al., 1985a; b; Yu, unpublished), the single C4B gene, C4B 3, present is therefore long. Four C4B clones were characterised from the AW genomic library constructed by K.T. Belt. Fine mapping of the cloned DNA confirmed that the C4B 3 gene possesses the 6-7 kb intron and is 22 kb in length.

In the individual AD (C4A 1; C4B Q0/ C4A 3; C4B Q0) only the 4.8 kb BamH I and 7.5 kb Kpn I fragments were detected by Southern blot analysis (Fig. 4.2, lanes 3 and 6). Therefore, the C4 genes present are all long genes. It has been suggested

Legend to Figure 4.1

Restriction maps of (a) long, (b) short human C4 genes and (c) the polymorphic C4d region.

- a. & b. Restriction maps of a long C4 gene (22kb) were derived from Cos 3A3 and that of a short C4 gene from λ JM-2a. Open dotted box indicates the approx. position of the 6-7kb intron. P_A and P_B are the 5' specific C4 cDNA probe and C4d specific genomic probe, respectively. For each map, the top represents the order and size of Bam HI fragments, and the bottom the Kpn I fragments. When P_A and P_B are used, long C4 genes have 4.8kb Bam HI and 7.5kb Kpn I fragments, and short C4 genes have 3.3kb Bam HI and 8.5kb Kpn I fragments.
- c. The 2.3kb Sma I fragment encodes >90% of C4d. Sm = Sma I, B = Bam HI, Nc = Nco I, X = Xho I, P = Pvu II, Na = Nar I (present in some C4 genes only; refer to text), and Hf^X = Hinf I (only the single site of the 927bp Bam HI fragment is shown). $\rightarrow$ depicts the position in the sequencing primer near to the isotypic region.



Legend to Figure 4.2

Genomic Southern analysis of AW, JM and AD DNA.

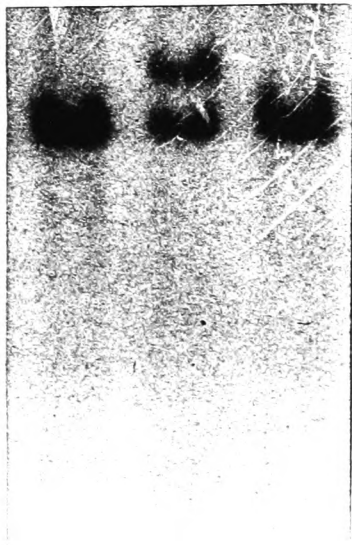
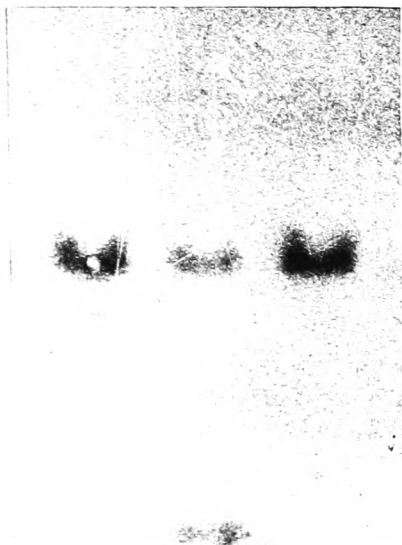
Lanes 1 and 4, AW (C4A 3; C4B 3/C4A 3; C4B 00); lanes 2 and 5, JM (C4A 4; C4B 5/C4A 3; C4B Q0); lanes 3 and 6, AD (C4A 1; C4B Q0/C4A 3; C4B Q0). Lanes 1-3 show DNA restricted with Bam HI and probed with the 5' specific probe, P_A. Lanes 4-6 show DNA restricted with Kpn I and probed with the C4d specific probe P_B. Long C4 genes are characterised by 4.8kb Bam HI and 7.5kb Kpn I fragments. Short C4 genes are characterised by 3.3kb Bam HI and 8.5kb Kpn I fragments.

Bam HI

Kpn I

1 2 3

4 5 6



— 8.5
— 7.5
— 4.8
— 3.3

the C4B Q0 allele of the HLA haplotype A3 B47 DR7 C4A 1; C4B Q0 is deleted (White et al., 1985; Carroll et al., 1985a; Schneider et al., 1986). On the other hand, the C4B Q0 allele of the HLA haplotype A2 B44 DR6 C4A 3; C4B Q0 is a long C4 gene and has been cloned (Chapter 6; Yu and others, in preparation).

Four C4A genomic clones were isolated and characterised from the AD library. In order to differentiate between the C4A 3 and the C4A 1 alleles a Bam HI/ Nar I double digest of the cloned DNA was carried out. This digest was chosen because all C4A alleles so far characterised contained a single Nar I site in the 927 bp Bam HI fragment (Fig. 4.1c) whereas the C4B alleles did not. It was anticipated that as C4A 1 had been serologically typed as Ch positive (i.e. contained the C4B antigenic determinants, Rittner et al., 1984) the C4A 1 allele may also not contain the Nar I site. The 927 bp BamH I fragment at the C4d region of three of the clones was cleaved to two 810 bp and 117 bp fragments by Nar I. The nucleotide sequence of the BamH I fragment from one of these clones, λ AD 2 α , was found to be identical to the sequence of the gene in Cos 3A3 previously assigned as coding for C4A 3a (Belt et al., 1985). (These three clones therefore correspond to C4A 3.) The Bam HI fragment in the fourth clone, λ AD-1 α , remained intact after the Nar I digest, and therefore correspond to the C4A 1 allele.

In the individual JM (C4A 4; C4B 5/C4A 3; C4B Q0) both 22 kb and 16 kb C4 genes were detected, as shown by the

presence of the 4.8 kb and 3.3 kb BamH I fragments (Fig. 4.2, lane 2) as well as the 7.5 kb and 8.5 kb Kpn I fragments (Fig. 4.2, lane 5) in Southern blot analysis. Further genomic Southern blot analysis shown that there were no long C4B genes present in individual JM and the polymorphic C4d region of the short C4 genes appeared identical in many marker restriction digests (data not shown). Two C4B clones were characterised from the JM library. One of these clones, λ JM-2a, includes a >17 kb insert that encodes a complete 16 kb C4B gene. The other clone, λ JM-6a, starts at the middle of the C4B gene and extends 3' to the 21-hydroxylase B gene. Fine restriction mapping showed that the polymorphic C4d regions of the two clones are identical. This implies that the C4B Q0 allele of JM is either deleted or is a short C4B gene which is identical to the C4B 5 allele at the C4d polymorphic region. Further studies favour the proposal that the C4B Q0 allele of the HLA A3 B35 haplotype is deleted (data not shown). The clones λ JM-2a and λ JM-6a are therefore assigned as coding for the C4B 5 allotype.

4.2.2 Sequence analysis of 9 allotypes of C4 at the C4d region

Comparison of nucleotide sequences from a full length and two partial C4 cDNA clones corresponding to C4A 4, C4A 3, and C4B 2, and limited genomic sequences of the C4B 1a, C4B 1b, and C4B 2 alleles (Belt et al., 1984; 1985) had already shown that there are only a total of 15 nucleotides changes present in the coding sequence, which results in 13 amino acid differences.

When additional data from amino acid sequencing of C4 proteins is included (Chakravarti et al., 1987), 15 amino acid differences are apparent, 12 of which are clustered in the section of the chain C-terminal to the thiolester. This very small number of changes is apparently responsible for the substantial differences in the reactivity of the thiolester bond, and for the charge and antigenic variation in different allotypes. How these structural and functional differences were related to changes in amino acid sequence was not clear. In order to define the amino acid changes responsible for these differences, the primary structure of the C4d polymorphic region from four informative alleles were determined: C4A 3a, C4A 1, C4B 3 and C4B 5 (Table 4.1). Parallel functional studies had already shown that all allotypes of each class possess their own characteristic differential reactivity (Dodds et al., 1986). C4A 3a was deduced to be Rg: 1,2; Ch: -1,-2,-3,-4,-5,-6 and C4B 3 was Rg: -1,-2; Ch: 1,2,3,4,5,6. The structure of these two isotypes can therefore give an account to the amino acid sequences that are related to the Rg/Ch antigenic determinants in addition to those that determine the thiolester reactivity. C4A 1 is an exceptional C4A variant because it does not possess either of the Rg determinants but acquires four of the Chido determinants (Rittner et al., 1984; Giles, 1987). It was typed Rg: -1,-2; Ch: 1,-2,3,-4,5,6. C4B 5 is essentially the reverse of C4A 1. It was typed Ch: -1,-2,-3,4,-5,6 and was Rg positive (Roos et al., 1984; Hing et al., 1986; Giles et al., 1984; Giles, 1987). By comparing the structural data of C4A 1 and C4B 5 with C4A 3 and C4B 3, the amino acid sequences involved in determining the thiolester

Legend to Figure 4.3

Summary of differences in (a) nucleotide sequences and (b) derived amino acid residues at the C4d region for nine C4 alleles/allotypes.

- a) C4d exon-intron structure is shown, position and nature of DNA sequence variations are indicated. Exons are numbered 1 to 8 in open boxes. Varied nucleotides are marked by asterisks and their presence in various alleles tabulated. Coding sequences are in capital letters and non-coding sequences in small letters. Notice the mutually exclusive variation pattern of DNA coding for amino acids (aa) 1054, 1101-1106, 1157 and 1186-1191. The C4bp-factor I cleavage sites are encoded by the first and the eighth exons, marked by vertical strokes.

Δ = deletion, □ = DNA coding for thiolester residues, ND = not determined.

- (b) Derived amino acid changes from (a) are shown. In addition Rg/Ch types are included (those of B 1a and B 2 were from deduction only) (Giles et al., 1984; Giles, 1987). Complete C4d fragment is from amino acids 938 to 1317 (numbering is for the continuous sequence of pro-C4).

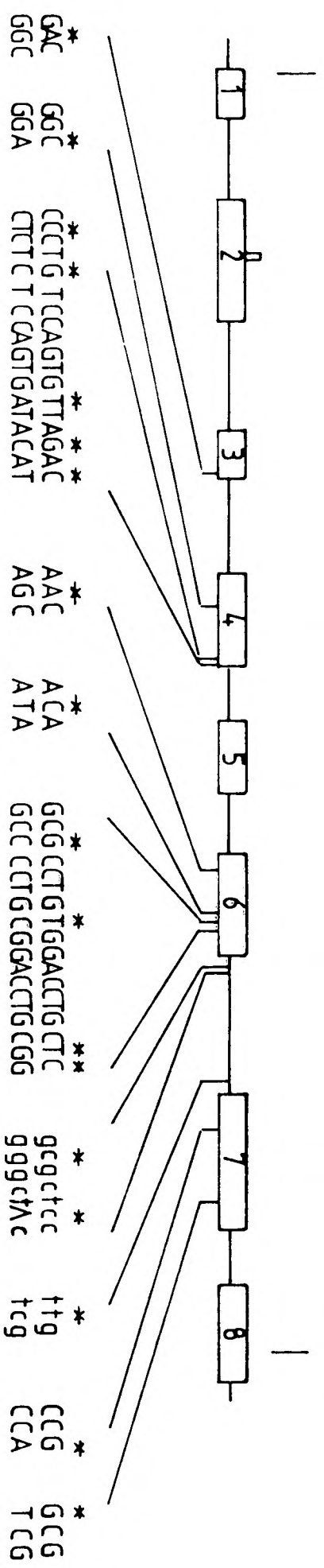
Residues in brackets (I and V) were found by amino acid sequencing, but were not detected in the DNA sequences.

Superscript c denotes data from cDNA sequence (Belt et al., 1984; 1985). Data on B 1a and B 1b were from Belt et al. (1985). □ = thiolester site; ▲ shows position of introns in genomic derived sequences.

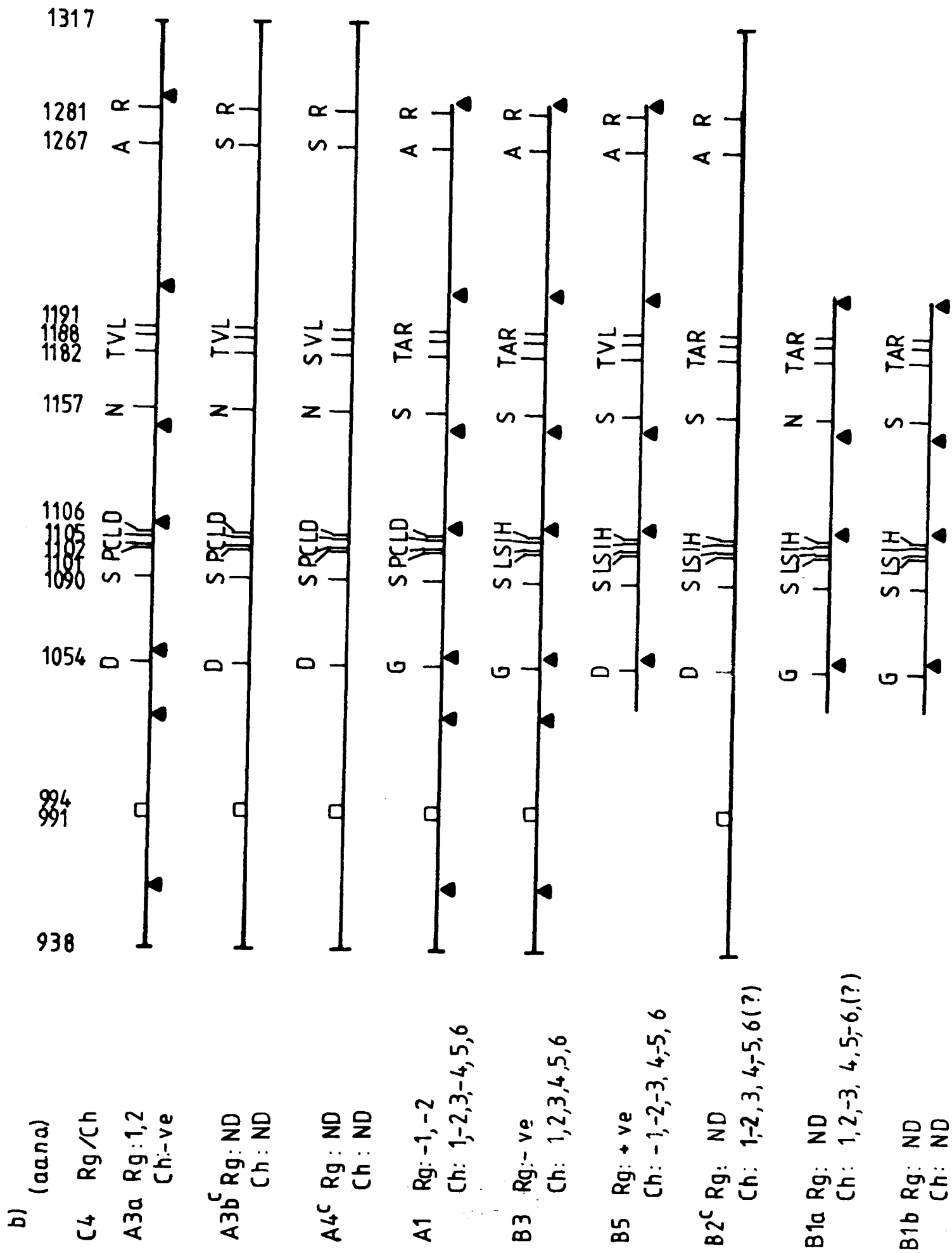
┌─┐ = boundary of C4d region, ND = not determined.

a)

0 100 200 bp



A3a	A	C	C	G	T	G	C	A	C	G	T	C	C	C	†	G	G
A3b ^c	A	C	C	G	T	G	C	A	C	G	T	TC	ND	ND	ND	G	T
A4 ^c	A	C	C	G	T	G	C	A	T	G	T	TC	ND	ND	ND	G	T
A1	G	A	C	G	T	G	C	G	C	C	C	GG	C	C	C	G	G
B3	G	C	C	T	A	C	T	G	C	C	C	GG	C	C	C	A	G
B5	A	C	C	T	A	C	T	G	C	C	G	TC	C	C	C	G	G
B2 ^c	A	C	C	T	A	C	T	G	C	C	C	GG	ND	ND	ND	G	G
B1a	G	C	C	T	A	C	T	G	C	C	C	GG	g	A	ND	ND	ND
B1b	G	C	C	T	A	C	T	G	C	C	C	GG	g	A	ND	ND	ND
aa No.	1054	1076	1101	1102	1105	1106	1157	1182	1186	1188	1191					1226	1267



(I)

(V)

reactivity and Rg/Ch antigenicity can be distinguished.

The variation of nucleotide sequences and their resulting amino acid changes, if any, of C4A 3a, C4B 3, C4A 1, C4B 5 and five other published alleles, C4A 3b, C4A 4, C4B 1a and C4B 2 (Belt et al., 1985) are depicted in Fig. 4.3a and Fig. 4.3b, respectively.

4.2.2.1 C4A 3a

As mentioned before, the C4A genes from Cos 3A3 and λ AD-2 α were identical and were assigned as C4A 3a. They differ by one nucleotide from a second C4A 3 allele, encoded in pAT-42 which is therefore assigned as C4A 3b. The C4A 3a allele in Cos 3A3 and λ AD-2 α encode an Ala residue at position 1267. However, due to a G to T transversion at the equivalent position in pAT-42, the C4A 3b allele encodes for a Ser residue instead (Belt et al., 1985).

The complete nucleotide sequence of the C4d region of the C4A 3a allele has been determined by the M13 sequencing method using randomly sonicated DNA. Comparison of the genomic DNA sequence with the cDNA sequence showed that the 380 amino acid C4d fragment is split by seven introns into eight exons that altogether span 2050 bp. The exons are termed D1 to D8 (which correspond to exons 24 to 31 in the C4 gene; chapter 3) and they encode 30, 70, 26, 52, 39, 57, 78 and 56 amino acid residues, respectively (Fig. 4.3 and Fig. 4.5b). The C4bp-factor I cleavage sites are encoded in exon D1 (24) and D8 (31), respectively, as shown in Fig. 4.3a. The residues involved in

the thiolester bond are encoded near the middle of exon D2 (i.e. exon 25 of the C4 gene). With reference to the previous results (Belt et al., 1985), exons D1, D2, D5 and D8 appear highly conserved and no polymorphic site has been detected so far. More than 90% of the C4d coding sequence is encompassed in a unique 2.3 kb Sma I fragment. Within this fragment is a 927 bp BamH I and a neighbouring 410 bp BamH I - Sma I subfragment (Fig. 4.1c) which together cover all the reported polymorphic sites of the C4 chain.

4.2.2.2 C4B 3

The clone λ AW-3B encodes the C4B 3 allotype which was serologically typed as fully Rg negative and fully Ch positive. The complete DNA sequence of the 2.3 kb Sma I fragment was determined. Comparing C4B 3 with C4A 3a, a total of thirteen nucleotides changes was detected. These resulted in 8 amino acid changes at four regions, i.e. at positions 1054, 1101-6, 1157 and 1188-91, of which C4B 3 has the sequence G-LSPVIH-S-ADLR and C4A 3a has the sequence D-PCPVLD-N-VDLL. Since these eight amino acid residues are the only detected changes at the C4d region between the two typical C4 isotypes, they must be related to the thiolester reactivity and to the Rg/Ch antigenic determinants. These eight amino acid changes can also explain that, in general, C4A is slightly acidic and C4B is slightly basic. C4B is associated with two basic amino acid residues (His 1106 and Arg 1191) and C4A with two acidic residues (Asp 1054 and Asp 1191).

The C4B 3 sequence G-LSPVIH-S-ADLR is also common to

C4B 1b (Belt et al., 1985). The faster mobility of C4B 3 than C4B 1 on agarose typing gel (Mauff et al., 1983a) is therefore due to the presence of some other polymorphic sites on C4B 3, presumably outside the C4d region. Two other C4B allotypes, C4B 1a and C4B 2, differ from C4B 3 each at a single position. At position 1157, C4B 1a has the C4A 3-related Asn residues instead of the C4B 3-related Ser residue. At position 1054, C4B 2 has the other C4A 3-related Asp residue instead of the C4B 3-related Gly residue (Fig. 4.3b).

4.2.2.3 C4A 1

The C4A 1 allotype encoded in the clone λ AD-1 α was serologically typed as Rg: -1,-2 and Ch: 1,-2,3,-4,5,6. It is a very interesting C4A variant not only because it expressed four of the Chido determinants and neither of the Rodgers determinants generally associated with C4A, but it is also the slowest migrating C4A allotype in the agarose typing gel (Rittner et al., 1984; Roos et al., 1984; Mauff et al., 1983a). The DNA sequence of the 2.3 kb Sma I fragment of the C4A 1 allele from clone λ AD-1 α was completely determined. Derived amino acid sequence at the C4d region revealed that it is almost identical to that of the C4B 3 allotype except at position 1101-6. At this region, C4A 1 retains the C4A 3-related sequence, PCPVLD, but the four amino acid residues at the other three polymorphic sites are all C4B 3-like. In other words, the C4A 1 has G-PCPVLD-S-ADLR compared with D-PCPVLD-N-VDLL for C4A 3a, and G-LSPVIH-S-ADLR for C4B 3. Since C4A 1 is a typical C4A allotype with respect to its thiolester reactivity (Dodds et al., 1986), this suggests that the four

amino acid differences at position 1101-6, i.e. PC--LD in C4A versus LS--IH in C4B, are solely responsible for the differences in the thiolester reactivity observed between the C4A and C4B isotypes. Compared with C4B 3, the C4A 1 allotype does not possess the Ch2 and Ch4 antigenic determinants. This infers that the C4B isotypic amino acid residues are involved in expressing the Ch2 and Ch4 epitopes because they are the only difference between C4A 1 and C4B 3 at the C4d region. However, the C4A isotypic residues, PC--LD, are not related to either of the two Rg determinants because C4A 1 is Rg negative. The amino acid variations at the other three polymorphic regions, i.e. at position 1054, 1157 and 1188-91, also appear related to the expression of Rg/Ch antigenic determinants.

4.2.2.4 C4B 5

Historically, C4B 5 was the first C4B allotype that was demonstrated as having both Rg and Ch determinants expressed on the same molecule (Roos et al., 1984). To differentiate C4B 5 from another allotype with the same electrophoretic mobility, this variant was also designated as C4B 5Rg⁺. The C4B 5 allotype encoded in clone λJM-2a was serologically typed as Rg positive and Ch: -1,-2,-3,4,-5,6. The DNA sequence of the 927 bp BamH I fragment and its neighbouring 410 bp BamH I - Sma I fragment were determined. The derived amino acid sequence at the C4d polymorphic region reveals that the C4B 5 allotype retains the C4B 3-related sequences at two regions, 1101-6 and 1157. However, the sequences at the other two regions, 1054 and 1188-91 are both C4A 3-like, i.e. C4B 5 has the sequence D-

LSPVIH-S-VDLL. This again suggests that the amino acid sequence at position 1101-6, LS--IH, is C4B isotypic and is involved in determining the C4B class specific properties. The C4B 5 allotype expresses Ch4 and Ch6. Its two C4B 3-related sequences, LSPVIH 1101-6, and S 1157, are therefore related to these two Chido antigenic determinants.

4.2.2.5 Comparison of intron sequences of C4d

One of the very striking features of the DNA polymorphism of C4 alleles is that their intron sequences at the C4d region are highly conserved. From the data available, the seven introns in the 2.3 kb Sma I fragment of C4B 3 and C4A 1 are totally identical, and the C4A 3a differs only by a c to t transition in the intron 3' to exon D6 (see Fig. 4.3a). In addition, the sequences of the five introns of the C4B 5 allele are identical to those in C4B 3. This reinforces the previous report of Belt et al., (1985), which compared the introns in the 927 bp BamH I fragments among the C4A 3a, C4B 2, C4B 1a and C4B 1b alleles. Two differences in the intron 3' to exon D6 of the C4B 2, C4B 1a and C4B 1a alleles compared with other alleles sequenced have been defined. One is a g to c transversion. The second is the deletion of a c nucleotide (Fig. 4.3a and Belt et al., 1985).

4.3 Discussion

In this chapter, I described four findings on the genetics and polymorphism of the human complement components C4A and

C4B: (1) C4B genes of different gene size (i.e. 22 kb vs. 16 kb) were cloned, (2) C4A and C4B can be defined by four isotypic amino acid residues at position 1101-6, which are likely to be responsible for their different thiolester reactivities; (3) the probable locations of the two Rg and six Ch antigenic determinants have been deduced and (4) the pattern of DNA polymorphism of C4A and C4B at the C4d region seems to be the result of gene conversion events. For this study, three genomic libraries were constructed and four interesting C4 genes representing two alleles from each class were isolated and partially characterised.

Variation in the number of expressed C4 gene has been noted for ten years (Teisberg et al., 1976), but differences in the gene size at the same locus is not a common phenomenon. It is now clear that the C4B genes can be either 22 kb or 16 kb in size, as exemplified by the cloned C4B 3 and C4B 5 alleles, respectively. Southern blot analysis has also indicated that a portion of the C4B 1 alleles (such as those linked to C4A 3) are also 22 kb in size (Palsdottir et al., 1987). Three other C4B genes from cosmid clones coding for C4B 2, C4B 1a and C4B 1b, respectively, were found to be 16 kb in size only (Carroll et al., 1985a). No 16 kb C4 A gene in the first C4 locus (see chapter 6) has been detected so far. Detailed restriction mapping and Southern blot analysis suggested that a long C4 gene is due to large intron lying about 2-2.5 kb 3' to the transcription start site. The presence of the very similar C4A and C4B genes in tandem might have caused some mispairing of the non-allelic genes between homologous chromosomes during

meiosis with unequal crossing over (Carroll et al., 1985b). The widespread distribution of two homologous C4B genes with a size difference of about 6-7 kb would increase the chance of mispairing between the long C4B and the long C4A gene. This is of particular interest because of the very high frequency (~10-15%) of the C4A or C4B null alleles in the population, and deletion or duplication of the C4 genes is not unusual (Carroll et al., 1985b; Hauptmann et al., 1986). Obviously, it is important to determine the nature of the 6-7 kb intron and to see whether it is involved in influencing the recombinational events of C4 genes and their extensive polymorphism. Preliminary data show that this 6-7 kb intron could be a member of the long interspersed sequences, LINE or L1 (Singer, 1982).

The structural data on the complete Rg⁺ C4A 3a and the complete Ch⁺ C4B 3 show that there are eight amino acid changes at four discrete regions between a common C4A and a common C4B allotype. At position 1054 (I), 1101-6 (II), 1157 (III) and 1188-91 (IV), C4A 3a is characterised by the sequence D-PCPVLD-N-VDLL while C4B 3 is characterised by the sequence G-LSPVIH-S-ADLR (Fig. 4.4). These are the only detectable changes in the polymorphic C4d region and therefore account for the many differences between the C4A and C4B isotypes. These included the differential reactivities of the labile thiolester, the apparent α chain molecular weight difference of 2,000 daltons, and the Rg/Ch antigenic determinants.

Structural information from other less common variants of

each isotype, C4A 1 and C4B 5, contributes to the understanding of the molecular basis of the isotypic properties and on the location and nature of the Rg/Ch antigenic determinants. The C4A 1 and C4B 5 proteins possess their class-specific properties, but essentially the reversed antigenicities. C4A 1 is characterised by the sequence G-PCPVLD-S-ADLR and C4B 5 is characterised by the sequence D-PCPVLD-S-VDLL. A combination of the structural data on C4A 3a, C4A 1, C4B 3, C4B 5 and five other published C4 allotypes (Belt et al., 1985) conclusively suggests that the amino acid changes at regions (I), (II), and (IV) are only related to the Rg/Ch antigenic differences and there are only four amino acid changes common to all C4A and to all C4B allotypes (Fig. 4.4). The C4A isotype can be defined by PCPVLD, and C4B defined by LSPVIH at position 1101-6 (i.e. region II). Apparently, these four isotypic residues are the root of the many class specific properties. The previous suggestion that the VDLL/ADLR at 1188-91 (Belt et al., 1985; Law et al., 1984a) might be C4A/C4B isotypic residues can now be ruled out.

Thus it appears that the C4 isotypic region is responsible in modulating the thiolester reactivity. The thiolester and the isotypic residues are separated by 106 amino acids. A hydropathy plot of C4d (Fig. 4.5a) shows that the isotypic residues are located at a hydrophilic region separated from the thiolester by two hydrophobic zones. The thiolester residues appear at the middle of a hydrophobic pocket. It is likely that the proposed conformational change during C4 activation by $C1$ (Iseñan and Kells, 1982) would bring the isotypic and the

Legend to Figure 4.4

Isotypic residues and possible location of Rg/Ch antigenic determinants.

Isotypic residues of C4A and of C4B are marked by crosses. The probable location of Rg/Ch determinants are indicated by 'O'. The C4A isotypic residues are not related to the Rg determinants. The two Rg epitopes lie on two of the regions I, III and IV on C4A. The six Ch determinants are formed by conformational and probably sequential epitopes from the four polymorphic sites on C4B. ▲ shows position of introns in genomic derived sequences.

—S— = thiolester site.

thiolester residues close together for certain interactions. The negatively charged Asp 1106 of C4A might increase the nucleophilicity of the nearby amino or amide groups towards the thiolester carbonyl (Law et al., 1984a). This may provide an explanation for the higher covalent binding affinity of C4A to peptide antigens. By the same analogy, the positively charged His 1106 of C4B might participate in hydrogen-bonding charge shift in hydroxyl substrates, resulting in more efficient binding to hydroxyl groups of carbohydrate antigens. This postulation leads to one fundamental question on the mechanism of the breakage of the thiolester bond and its reactivity. Instead of the generally suggested conformational change exposing a reactive thiolester which undergoes rapid reaction with any available nucleophile, a situation may occur in which a conformational change rearranges the structure of the isotypic and the thiolester regions, forming a transient state that can catalyse the nucleophilic attack by the amino or hydroxyl groups on the thiolester carbonyl.

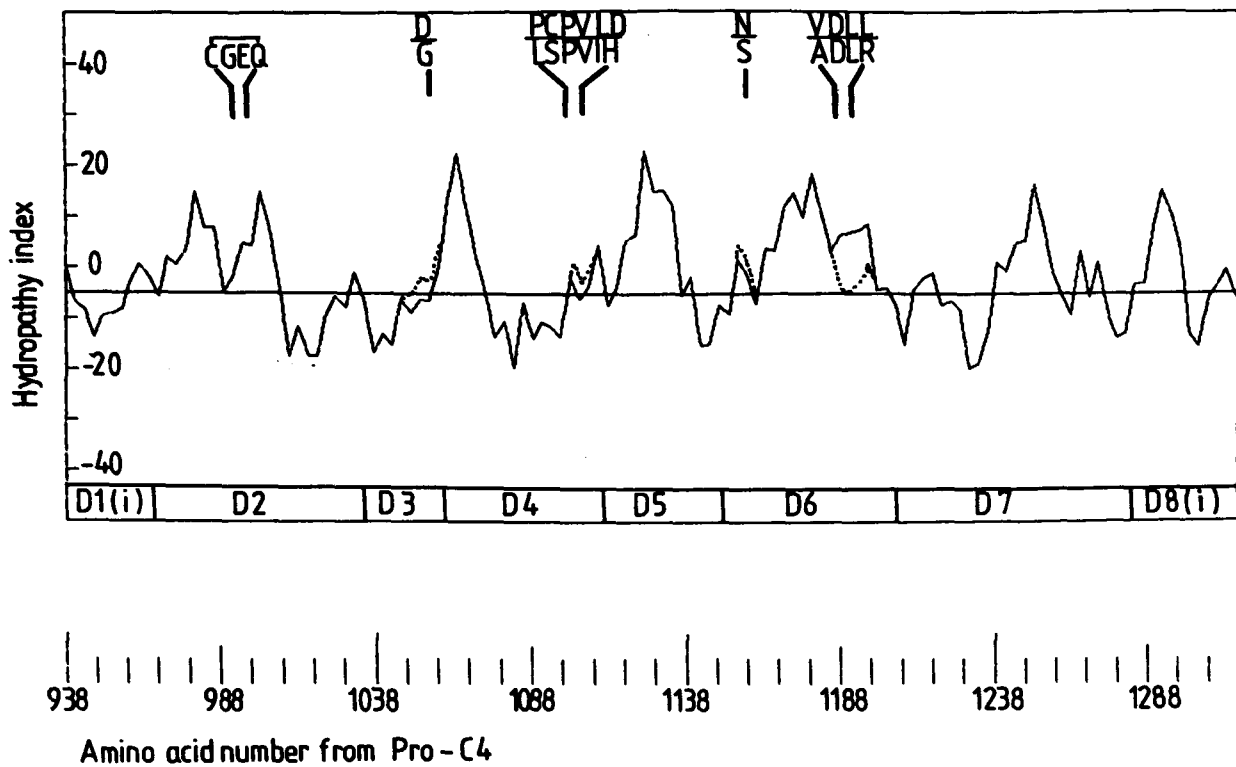
Besides the Asp/His 1106, the other three isotypic residues and the sequence nearby (Fig. 4.5b) might be important in modulating the thiolester reactivity as well. It is possible that the sulphhydryl group of Cys 1102 in many C4A may interact with the carboxyl of Asp 1106 or with the sulphhydryl of the thiolester Cys 991 at some stage. There is no other Cys residue at the C4d region of either isotype. The Pro 1101 of C4A would possibly cause considerable conformational change at the isotypic region. Together with Pro 1103, Pro 1101 may impose significant constraints on the local structure. It has

Legend to Figure 4.5

Comparison of (a) hydropathic profiles and (b) derived amino acid sequences of C4d fragments between complete Rg⁺ C4A and complete Ch⁺ C4B.

- (a) The computer program HYDROLOT, based upon the algorithm of Kyte and Doolittle (1982), was employed with a window of 11 amino acids. C4A profile is shown and the differences between C4A and C4B is in dotted lines. Position of the thiolester residues, and polymorphic residues between Rg⁺ C4A and Ch⁺ C4B are indicated. The region encoded in exons D1 to D8 is represented in boxes. The C4d boundaries are encoded by exons D1 and D8 (see text). (i) = incomplete exon.
- (b) The corresponding amino acid sequence for (a). The complete Rg⁺ C4A is based on data from C4A 3a. Complete Ch⁺ C4B is based on data from C4B 3 for sequence encoded by exons D1 to D7 and C4B 2 for that of exon D8. Positions of the exon boundaries are marked by vertical strokes. The thiolester site is overlined. The isotypic residues for C4A and C4B are marked by asterisks. The probable location of Rg/Ch determinants are marked by #, and * (for C4B).

a)



b)

938

C4A TLEIPGNSDP NMIPDGFNS YVRVTASDPL DTLGSEGALS PGGVASLLRL PRGGEQTM

C4B TLEIPGNSDP NMIPDGFNS YVRVTASDPL DTLGSEGALS PGGVASLLRL PRGGEQTM

998

C4A YLAPTLAASR YLDKTEQWST LPPETKDHAV DLIQGYMRI QQFRKADGSY AAWLSRDSST

C4B YLAPTLAASR YLDKTEQWST LPPETKDHAV DLIQGYMRI QQFRKADGSY AAWLSRGSST

1058

C4A WLTAFLVKVL SLAQEQVGGG PEKLQETSNW LLSQQQADGS FQDPCPVLD R SMQGGVLGND

C4B WLTAFLVKVL SLAQEQVGGG PEKLQETSNW LLSQQQADGS FQDLSPVIH R SMQGGVLGND

1118

C4A ETVALTAFVT IALHHGLAVF QDEGAEPKQ RVEASISKAN SFLGEKASAG LLGAHAAIT

C4B ETVALTAFVT IALHHGLAVF QDEGAEPKQ RVEASISKAS SFLGEKASAG LLGAHAAIT

1178

C4A AYALTTLKAP VDILGVAHNN LMAMAQETGD NLYWGSVTGS QSNVSPTPA PRNPSPMPQ

C4B AYALTTLKAP ADLRGVAHNN LMAMAQETGD NLYWGSVTGS QSNVSPTPA PRNPSPMPQ

1238

C4A APALWIETTA YALLHLLHE GKAEMADQAS AWTROGSFQ GGFRSTQDTV IALDALSAYW

C4B APALWIETTA YALLHLLHE GKAEMADQAA AWTROGSFQ GGFRSTQDTV IALDALSAYW

1298

C4A IASHTTEERG LNVLSSTGR

C4B IASHTTEERG LNVLSSTGR

been suggested that Pro 1101 and Cys 1102 may be responsible for the anomalous slower mobility of the C4A α chain in the modified SDS-polyacrylamide gel (Chan and Atkinson, 1985).

There is only one functional C4 gene in mouse, which has to cope with both peptide and carbohydrate antigens. The mouse C4 protein appears to be a hybrid of human C4A and C4B, with respect to the isotypic residues, i.e. PCPVIH (Levi-Strauss et al., 1985; Sepich et al., 1985; Nonaka et al., 1985). The first two residues are human C4A specific, but the last two are C4B specific. It will be of interest to know if the reactivity of other thiolester-containing proteins such as the mouse C4, the complement component C3 and the protease inhibitor α_2 macroglobulin are also modulated by the area analogous to the human C4 isotypic region. A comparison of the amino acid sequences of these thiolester-containing proteins at the area analogous to the human C4 isotypic region is shown in Table 4.2. Since mouse C4, human C3 and mouse C3 all possess the human C4B specific sequence, IH at position 1105-6, they might have high covalent binding affinity for hydroxyl groups and this has indeed been shown for human C3 (Law et al., 1984b).

Analysis of the sequence and serological data among C4A 3a, C4A 1 and C4B 5 sheds light on the location of the six Ch antigenic determinants. There is not a direct one-to-one correlation between amino acid substitution at each of the four polymorphic regions and antigenic determinants. At the C4d region, C4A 1 has the identical amino acid sequence to that of C4B 3 except at the isotypic residues and C4A 1 expresses all

TABLE 4.2 Comparison of amino acid sequences at the thiolester sites and at the region analogous to the human C4 isotypic site of some thiolester-containing proteins.

	Thiolester site	Region analogous to the human C4 isotypic site	References
Human C4A Human C4B	PRGCGEQTMI PRGCGEQTMI	^{**} [*] FQDP [*] CPVLDRSM FQDLSPVIHRSM	Campbell et al., 1981; Hellman et al., 1984; Belt et al., 1984; 1985; this work.
Murine C4	PQGCAEQTMI	FHDPCPVIHRAM	Levi-Strauss et al., 1985; Sepich et al., 1985; Nonaka et al., 1985.
Murine S1p ⁺	PRSCAEQTMI	FHDPCPVIHRAM	Ogata and Sepich, 1985; Nonaka et al., 1986.
Human C3 Murine C3	PSGCGEQNMI PAGCGEQNMI	FQEDAPVIHQEM FQEDGPVIHQEM	De Bruijn and Fey, 1985. Wetzel et al., 1984.
Human α_2^M	PYGCGEQNMV	FRSSGSLNNAI	Kan et al., 1985.

(⁺ haemolytically inactive)

Ch determinants except Ch2 and Ch4. This suggests the C4B isotypic residues are involved in the expression of these two epitopes. The conversion of G to D at position 1054, and ADLR to VDLL at positions 1188-91 result in the loss of the Ch 1, 2, 3, and 5 epitopes on the C4B 5 molecule. Therefore, the C4B 3-related G at position 1054 and ADLR at 1188-91 seem related to Ch 1, 2, 3 and 5. It is apparent that the structural component of each of the four discrete polymorphic sites is involved in the formation of more than one epitopes. On the other hand, the presence of G 1054, S 1157 and ADLR 1188-1191 on C4A 1 only forms Ch 1, 3, 5, 6, and the presence of LSPVIH 1101-6 and S 1157 on C4B 5 only forms Ch4 and Ch6. This would suggest that the six Ch antigenic determinants are formed by overlapping epitopes, some of which are conformational (Benjamin et al., 1984). It seems each of the four discrete polymorphic sites does not only form a possibly sequential epitope, but will also interact with the structural component of another epitope to give an additional conformational antigenic determinant. To this end, Ch4 and Ch6 could be sequential epitopes and Ch2 a conformational epitope. Using a similar approach, the probable location of the two Rg determinants can be derived by comparing sequence and serological data between C4A 3a and C4A 1. They may fall within D 1054, N 1157 and VDLL 1188-91, although their exact correlation is yet unclear. Contrary to the situation in C4B, the C4A isotypic residues seem non-immunogenic and are not related to either of the two Rg determinants, as exhibited by the C4A 1 allotype. This would again reflect the conformational constraints imposed by the C4A isotypic amino acid residues. A structural model for the exact location of the

two Rg and six Ch determinants has been constructed and described in chapter 6.

The structural studies on 9 alleles from C4A or C4B show that there is a mosaic polymorphic pattern which is due to the shuffling of four discrete regions, or makers, within 1 kb of DNA at the C4d region. This is also manifested by the mixing of the two generally discrete phenotypes, i.e. C4A with two Rg and C4B with six Ch determinants. The interchange of the discrete sequences at region I, III and particularly IV, will result in the expression of the reversed antigenicity. Of equal possibility, though not shown here, the interchange of the sequence at region II, i.e. the isotypic region, will convert the C4 protein expressed to the opposite class (chapter 6). On the other hand, it is interesting to note that although there are some thirteen changes in nucleotides at the coding region between the two isotypes at C4d there are only one or two changes in the intron sequences. Moreover, the overall G/C content for the 2050 bp of DNA coding for C4d is relatively high, i.e. 60%. One of the most plausible explanations for this mosaic polymorphism pattern is by some gene conversion-like events (Petes and Fink, 1982) between the repeated C4A and C4B genes. The high G/C content at the C4d region would favour this event to occur (Weiss et al., 1983a; 1983b). Alternatively, it could be caused by many independent, double crossing over events but this seems unlikely. As discussed before, it will be very interesting to determine whether the extensive C4 polymorphism and the unusually high frequency of recombination events are correlated. Knowledge of this would be helpful to

our understanding of the recombination mechanisms in human genetics.

4.4 Summary

The human complement component C4A and C4B are highly homologous proteins, but they show markedly different, class-specific, chemical reactivities. They also differ serologically in that C4A generally expresses the Rodgers (Rg) blood group antigens while C4B generally expresses the Chido (Ch) blood group antigens. C4A 1 and C4B 5 are exceptional variants which possess their class-specific chemical reactivities, but express essentially the reversed antigenicities. The genes encoding the typical Rg positive C4A 3a and Ch positive C4B 3 allotypes and the interesting variants C4A 1 and C4B 5 have been cloned. Characterisation of the cloned DNA has revealed that the genes encoding the C4A 3a, C4A 1 and C4B 3 allotypes are 22 kb long, but that encoding C4B 5 is only 16 kb long. Comparison of the derived amino acid sequences of the polymorphic C4d fragments has shown that C4A and C4B can be defined by only four isotypic amino acid differences at position 1101-6. Over this region, C4A has the sequence PCPVLD while C4B has the sequence LSPVIH, and this presumably is the cause of their different chemical reactivities. Moreover, the probable locations of the two Rg and the six Ch antigenic determinants have been deduced. The structural data on the C4A and C4B polymorphism pattern suggests a gene conversion-like mechanism is operating in mixing the generally discrete serological phenotype between C4A and C4B.

CHAPTER FIVE

A STRUCTURAL MODEL FOR THE LOCATION OF THE RODGERS AND THE CHIDO ANTIGENIC DETERMINANTS AND THEIR CORRELATION WITH THE HUMAN COMPLEMENT C4A/C4B ISOTYPES (A Review)

5.1 Rodgers and Chido are C4 antigens

The Chido (Ch) and Rodgers (Rg) are human plasma and red blood cell antigens which were defined by the alloantibodies, anti-Ch (Harris et al., 1967; Middleton and Crookston, 1972) and anti-Rg (Longster and Giles, 1976). Anti-Rg and anti-Ch are IgG antibodies formed by transfused patients who lack either Rg or Ch antigen, respectively. The genes coding for the Rg and Ch antigens were both mapped to the human leucocyte antigen system, HLA (Middleton et al., 1974; Giles et al., 1976). The structural genes coding for the fourth component of complement C4 (Rittner et al., 1975; Teisberg et al., 1976), and the complement components C2 (Fu et al., 1974) and factor B (Allen, 1974), were also mapped to the HLA region. Biochemical, genetic and serological studies revealed that Rg and Ch are distinct antigenic components of the two isotypes of C4, C4A and C4B (previously called C4F and C4S), respectively (O'Neill et al., 1978a, b). Subsequently it was shown that the Rg and Ch antigens were both located in the C4d regions of C4A or C4B (Tilley et al., 1978). C4d is the factor I mediated proteolytic degradation fragment of C4 that may be covalently linked to various cell surfaces (see Reid and Porter, 1981 for a review). C4A and C4B are highly homologous in structure (Belt et al., 1984, 1985; Yu et al., 1986), but possess markedly different

electrophoretic mobilities, haemolytic activities (see Mauff et al., 1983a, for a review), and chemical reactivities (Isenman and Young, 1984, 1986; Law et al., 1984a; Dodds et al., 1986; Shrifferli et al., 1987). The latter is due to the differential reactivities of the thiolester carbonyl group in the two classes of C4: C4A exhibits relatively higher covalent binding affinity to amino groups or peptide antigens than C4B, while C4B binds preferentially to hydroxyl groups or carbohydrate antigens. This phenomenon may also explain the higher haemolytic activity of C4B in the conventional haemolytic assay that is performed using sensitized sheep erythrocytes whose surface is rich in carbohydrate antigens (Table 5.1). The identification of C4A and C4B is based functionally on the haemolytic activity and covalent binding activity, or structurally on the apparent mol. wt. of the alpha chains in modified SDS-polyacrylamide gel electrophoresis (C4A α - 96 000; C4B β - 94 000; Roos et al., 1982), or serologically using anti-Rg and anti-Ch antisera.

5.2 Polyspecificity of the alloantisera and multiple determinants for Rg and Ch

The haemagglutination inhibition test that was initially developed to study markers or allotypes of immunoglobulin, Gm (Grubb, 1956, reviewed in Nisonoff et al., 1975) has been widely applied in the serological typing for the Rg and Ch antigens (Middleton and Crookston, 1972; Giles, 1980). To test the antigenicity of C4, the plasma is examined for its ability to inhibit agglutination in an anti-globulin test: (1) the plasma (antigen) is mixed with anti-Rg or anti-Ch (antibody) to allow Ab-Ag interaction and (2) subsequently added to human

Table 5.1 Comparison of human C4A and C4B protein

	C4A	C4B
1. Electrophoretic mobility		
a) agarose gel	Fast (acidic)	Slow (basic)
b) SDS-PAGE (α chain)	96K	94K
2. Thiolester reactivity		
a) Haemolytic activity	Lower	Higher
b) Relative covalent binding affinities		
i) amino group	Higher	Lower
ii) hydroxyl group	Lower	Higher
3. Antigenic determinants [*]	Rodgers	Chido
(Blood group antigen)	(Rg: 1,2)	(Ch: 1,2,3,4,5,6)

^{*} with exceptions such as C4A 1 and C4B 5

erythrocytes coated with standard C4 (free anti-Rg/anti-Ch would bind to the coated C4); the C4-coated red cells are then spun down, washed and tested for the ability to be agglutinated by anti-IgG. If the test plasma contains the corresponding antigens for anti-Rg or anti-Ch, its interaction with the antisera will eliminate the subsequent reaction of the alloantibody with the standard C4 on red cells, therefore, no anti-globulin agglutination follows. Thus the C4 molecules in the test plasma are termed 'complete inhibitors'. On the contrary, if the antigens in the test plasma do not match with the test anti-Rg or anti-Ch, the antibodies will bind to standard C4 coated on red cell. Visible agglutination among red cells will result subsequent to the addition of anti-IgG to the reaction mixture. Thus the C4 molecules in the test plasma are called 'negative inhibitors' (Giles, 1980). On the other hand the specificities of the antisera can be examined by using plasma of known Rg/Ch type in the test. As an alternative to this approach the test C4 can be coated on red cells and then examined for its ability to be agglutinated by anti-Rg or anti- Ch.

In a study on the anti-Rg, Longster and Giles (1976) observed that some plasma samples exhibit a partial inhibition (p.i.) phenomenon. In other words the Rg/Ch antigens from some individuals show weaker interaction with certain anti-Rg or anti-Ch, although there is no indication that the related plasma C4 is expressed at lower level. p.i. was demonstrated for Ch in the C4B 2 allotype that is generally expressed in parallel with C4A 4 (Nordhagen et al., 1980). The p.i. phenomenon was also detected on the C4B 1 allotype that is generally expressed in parallel

with C4A 6 or to a proportion of C4A 3. However, it was found that the anti-Ch attributed to the p.i. in C4B 2 did not result in the p.i. phenomenon observed in the C4B 1, and vice versa (Giles, 1984). This was taken to suggest that the anti-Ch and anti-Rg are polyspecific and that there may be more than one antigenic determinant for Rg or for Ch. Different sources of antisera vary in antibody specificities for Rg/Ch determinants (Giles, 1985a). A single partial inhibitor pattern has been detected in Rg so far and is generally detected in the isoexpressed haplotype (see chapter 6 for definition) C4A 3 A 2. Thus Giles (1985b) proposed a model with two Rg and three Ch antigenic determinants and they are designated as Rg 1, 2 and Ch 1, 2, 3, respectively. In both cases, Rg1 and Ch1 were suggested to be the commonest epitopes.

Standardised polyspecific anti-Rg and anti-Ch alloantisera allowed serological characterisation of many rare C4 allotypes. This not only led to the subdivision of many common allotypes, but also the invalidation of the C4A-Rg and C4B-Ch dogma. Rare variants of both isotypes such as C4A 1 and C4B 5 which both possess their isotypic properties (i.e. haemolytic activity, covalent binding affinity to small molecules, and the apparent mol. wt. of the alpha chain) were found essentially expressing the reversed antigenic determinants (Rittner et al., 1984; Roos et al., 1984). Detailed analysis of the Rg positive C4B 5 protein (i) in the plasma for haemagglutination inhibition test, and (ii) coated on standard erythrocytes for direct agglutination experiment uncovered the presence of another 3 specificities of the anti-Ch and their corresponding antigenic determinants on C4 were designated Ch 4, 5 and 6 (Giles, 1987). Correlation of these

determinants with the Ch 1, 2 and 3 and with the Rg 1 and 2 determinants was made among 131 donors (Giles, 1987).

Assignment of the presence of Rg/Ch determinants on each C4 allotype is usually complicated by the presence of more than one C4A or C4B gene in each individual and therefore family studies are always necessary. The expression of some informative Rg/Ch haplotypes from the C4A and C4B haplotypes are listed in Table 5.2. Cumulative serological observation (Giles et al., 1984; Giles, 1987) enabled the following conclusions to be drawn.

1. All C4B molecules express Ch4.
2. All C4B molecules expressing Ch 1, 2, 3 also express Ch 4, 5 and 6 but not Rg 1, 2.
3. Ch2 and Ch5 are closely related and never split on C4B molecule.
4. Ch3 appears related to Ch1 and Ch6 on both C4A and C4B.
5. Both Rg1 and Rg2 may not be expressed by some C4A variants. Nonexpression of Rg1 and Rg2 or both always renders the C4A protein possessing certain Ch epitopes, but never Ch2 and Ch4.
6. Rg1 and Ch1, Rg2 and Ch6 appear to be the alternative antigenic determinants on the C4 molecule.

As a result, it can be speculated that some of the Ch determinants might be overlapping with other determinants, and that Ch2 and Ch4 might be related to the region of the C4 protein that would determine its isotypic properties.

Table 5.2 Interrelationship among C4A/C4B and Rg/Ch haplotypes

<u>Donors</u>	<u>C4 haplotypes</u>	<u>Rg/Ch haplotypes</u>	<u>Ref.</u>
General cases			
F.P.	A3B1	Rg: 1,2 ; Ch: 1,2,3,4,5,6	1
AW*	A3B3	Rg: 1,2 ; Ch: 1,2,3,4,5,6	2,3
S.L.	AQOB1	Rg: -1,-2 ; Ch: 1,2,3,4,5,6	1
J.Mah	A3BQ0	Rg: 1,2 ; Ch: -1,-2,-3,-4,-5,-6	1
C.G.	A4B2	Rg: 1,2 ; Ch: 1,-2,3,4,-5,6	1
Str.	A4B5	Rg: 1,2 ; Ch: 1,-2,3,4,-5,6	1,4
Rare cases:			
M.C.	A3B1	Rg: 1,2 ; Ch: 1,2,-3,4,5,-6	1
D.W.	A 3 A2BQ0	Rg: 1,2 ; Ch: -1,-2,-3,-4,5,+6	1
D.C.	A3BQ0	Rg: 1,-2 ; Ch: -1,-2,-3,-4,-5,6	1
MS74	A2B3	Rg: N.D. ; Ch: -1,2,-3,4,5,-6	1,5
J.M.*	A4B5	Rg: 1,2 ; Ch: -1,-2,-3,4,-5,+6	2,4,6
A.D.*	A1BQ0	Rg: -1,-2 ; Ch: 1,-2,3,-4,5,6	2,7

* Genomic libraries have been made for these individuals

N.D. not determined

- Ref. 1. Giles, 1987
 2. Yu et al., 1986
 3. Dodds et al., 1986
 4. Hing et al., 1986
 5. Skanes et al., 1985
 6. Roos et al., 1984
 7. Rittner et al., 1984

5.3 Structural model for C4 isotypic and antigenic determinants.

In parallel with the advance in serological typing of C4, remarkable progress was made in the structural studies of C4 between 1981 to 1986. This was initiated with the determination of the amino acid sequence around the putative thiolester site and most of the C4d region (Campbell et al., 1981; Harrison et al., 1981; Chakravarti et al., 1983; Hellman et al., 1984), and subsequently continued by the isolation and DNA sequencing of C4 cDNA and genomic clones (Carroll and Porter, 1983; Carroll et al., 1984a, b; Belt et al., 1984, 1985). Meanwhile, disulphide linkages (Janatova, 1986; Seya et al., 1986b, and sites of glycosylation (Chan and Atkinson, 1985) and sulphation (Karp, 1984; Hortin et al., 1986b) of C4 have been elucidated. No difference has yet been found in these post-translational modifications between C4A and C4B. The differential isotypic properties as well as the complex antigenic determinants are therefore more likely due to by amino acid differences. Full length and partial cDNA (Belt et al., 1984, 1985) and genomic DNA sequences (Chapter 3), and partial protein sequences (Chakravarti et al., 1983) from pooled serum identified ~17 amino acid changes among some common C4A and C4B allotypes. Surprisingly, only a few variations lie on the beta and the gamma chains, but 12 changes clustered at a small region of 140 amino acids which is ~230 amino acid residues C-terminal to the thiolester site at the C4d region. Presumably, these 17 amino acid variations are the basis of the (i) differential isotypic properties, (ii) antigenic determinants, and (iii) some allelic

variations. The small number of variations on the beta and gamma chains are probably responsible for allelic variations or for some other undefined features as these changes are not common to either C4A or to C4B isotypes (Chapter 3).

Dissection of the isotypic and antigenic determinants was achieved by molecular cloning and DNA sequencing of many alleles of the C4 molecules (Yu et al., 1986) with defined functional (Dodds et al., 1986) and antigenic properties (Giles, 1987) (Table 5.3). Comparison of derived amino acid sequences at the C4d region between the Rg positive C4A 3a and the Ch positive C4B 3 suggested that there are eight amino acid changes which may be correlated with the isotypic and antigenic determinants. These changes are located at four regions as shown in Fig. 5.1. Regions I and III are due to single amino acid substitutions, while region II involves four clustered changes and region IV involves two changes. Thus C4A 3a has the sequence D-PCPVLD-N-VDLL and C4B 3 has the sequence G-LSPVIH-S-ADLL at these regions. Since the sequence of the Rg negative C4A 1 and C4B 3 are identical at the C4d region except for the four changes at region II, this suggests that PCPVLD / LSPVIH at position 1101-6 are the isotypic residues for C4A / C4B. Further data from the C4B 5 allotype supports this finding as it maintains the C4B-specific sequence LSPVIH 1101-6, but acquires the C4A-related residues, D 1054 and VDLL 1088-91.

Since the Rg and Ch determinants are located in the C4d region (Tilley et al., 1978) and the eight substitutions at regions I to IV are the only changes detected among the Rg positive C4A 3a, Chido positive C4B 3, Rg negative C4A 1, and Rg

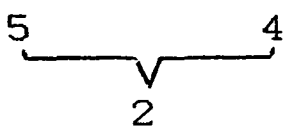
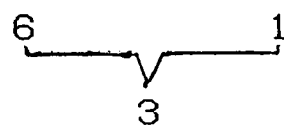
Table 5.3 Correlation of the serological and structural data in six cloned C4 genes.

<u>C4</u> <u>allotype</u>	<u>Rg / Ch haplotype</u>		<u>Characteristic</u> <u>a.a. sequence at C4d</u>			
			(I)	(II)	(III)	(IV)
A3a	Rg: 1, 2;	Ch: -1, -2, -3, -4, -5, -6	D	PCPVLD	N	VDLL
B3	Rg: -1, -2;	Ch: 1, 2, 3, 4, 5, 6	G	LSPVIH	S	ADLR
A1	Rg: -1, -2;	Ch: 1, -2, 3, -4, 5, 6	G	PCPVLD	S	ADLR
B5	Rg: 1, (ND);	Ch: -1, -2, -3, 4, -5, 6	D	LSPVIH	S	ADLR
B2	Rg: (ND);	Ch: 1, -2, 3, 4, -5, 6	D	LSPVIH	S	ADLR
B1a	Rg: (ND);	Ch: 1, 2, -3, 4, 5, -6	G	LSPVIH	N	ADLR

a.a. amino acid

positive C4B 5, these substitutions should also be related to the expression of the Rg/Ch antigenic determinants on the C4 molecules. Based on the structural data and the Rg and Ch haplotypes of these four cloned C4 genes, and also on the cDNA and genomic sequences of C4A 4, C4A 3b, C4B 1a, C4B 1b and C4B 2 whose Rg/Ch haplotype can only be derived indirectly (Table 5.3), it is possible to correlate some of the determinants with the amino acid residue changes at these four regions. For the Chido determinants, it appears that there is not a one-to-one direct correlation between an epitope and the amino acid changes(s) at each region. For examples, C4B 2 differs from C4B 3 only at region I (i.e. D 1054 cf. G 1054) and almost all C4B 2 is Ch-2 and Ch-5 (Giles et al., 1987); C4A 1 differs from C4B 3 at region II and is Ch -2 and Ch -4. Thus the amino acid change(s) at each region may be involved in the formation of more than one epitope. Some of the epitopes are discontinuous or conformational which involves more than one (or probably two) regions, this is in accord with serological results as already mentioned. Two C4B 1 genes were partially sequenced at the polymorphic C4d region (Belt et al., 1985). The sequence of C4B 1b is identical to that of C4B 3, while that of C4B 1a is unusual in having the C4A-related region IV (i.e. N 1157). It is possible that this C4B 1a corresponds to the 'partial inhibitor' which is Ch -3 and Ch -6. Based on these assumptions, the probable locations of the six Chido antigenic determinants have been deduced and are listed in Table 5.4. The C4B-related sequences at regions I, II, III and IV, respectively, constitute the Ch5, Ch4, Ch6 and Ch1 determinants. These four determinants may be continuous (or

Table 5.4 Comparison of Ch haplotypes and amino acid sequences of the C4d fragment of some C4 allotypes with a Ch positive C4B allotype (e.g. C4B 3)

<u>Polymorphic sequences at C4d</u>					
	(I) G 1054	(II) LSPVIH 1101-6	(III) S 1157	(IV) ADLR 1188-91	Ch haplotypes*
C4A 1		-			-2, -4
C4B 2	-				-2, -5
C4B 1a			-		-3, -6*
C4B 5	-			-	-1, -2, -3, -5
C4A 3	-	-	-	-	-1, -2, -3, -4, -5, -6
Deduced location of the Ch epitope					

- absence

* only negative haplotypes are shown

** assumed

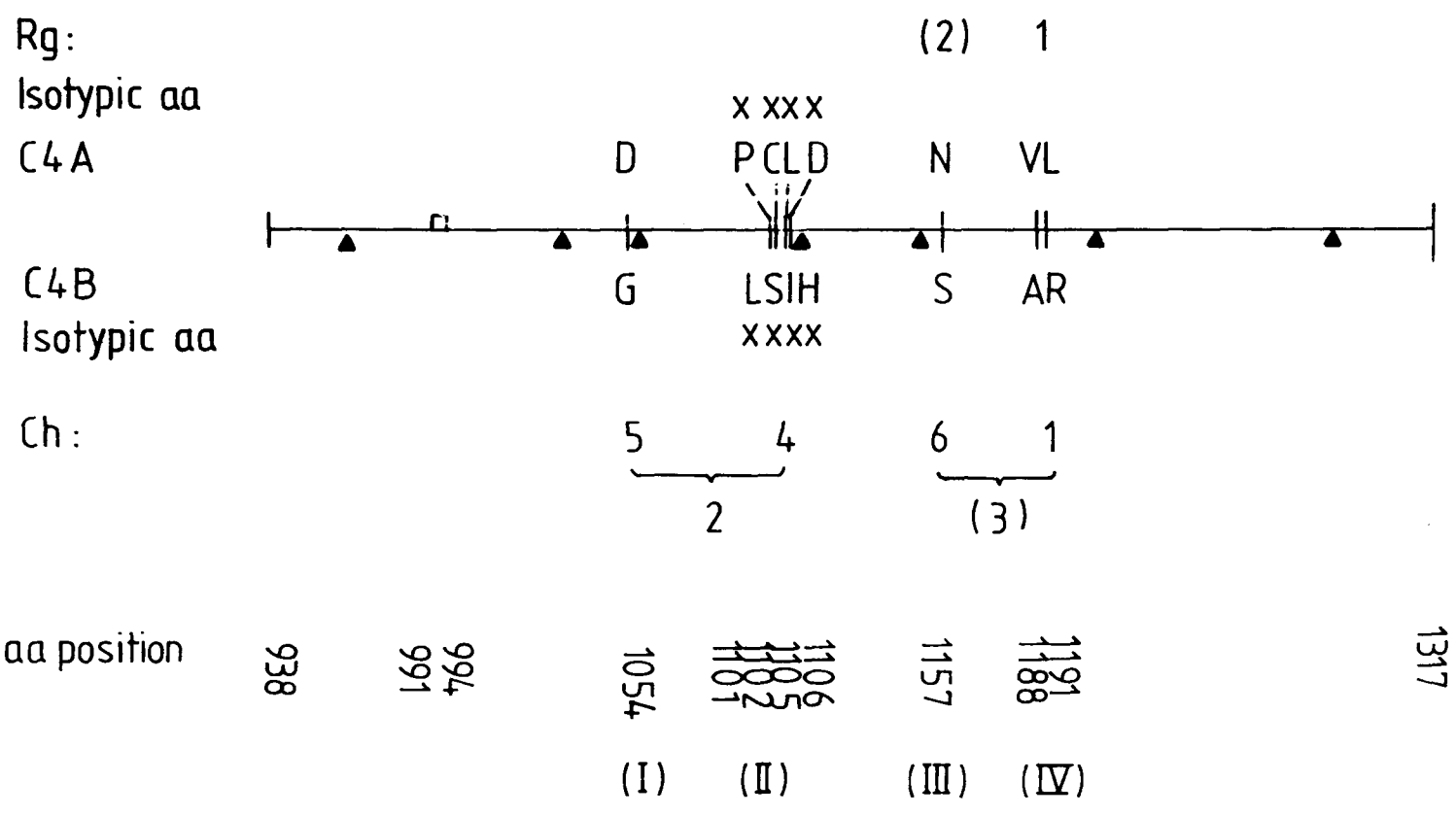
sequential) epitopes. On the other hand, Ch2 and Ch3 may be conformational epitopes: the former involves the amino acids from regions I and II, while the latter involves the amino acids from regions III and IV. As the C4B isotypic residues are related to the expression of Ch2 and Ch4, this explains the fact that these two epitopes are never expressed on the C4A molecules.

The situation for the location of Rg determinants is less obvious. As C4A 1 does not express Rg1 and Rg2, the C4A isotypic residues may not be related to the Rg determinants, unless they are involved in the formation of conformational epitopes. As C4B 5 expresses Rg1 and as Rg1 and Ch1 seem to be alternative in a C4 molecule, Rg1 is probably determined by the C4A-related sequence VDLL (1188-91). Accumulated serological data also suggest that Rg2 could be alternative to Ch6 (Giles, 1987), therefore, the Rg2 determinant might be located at the C4A-related region IV sequence (i.e. N 1157). Thus a molecular model for the location of the C4A/C4B isotypic residues and the Rg/Ch antigenic determinants has been formulated and is presented in Fig. 5.1.

This model can be partly tested by determining the DNA sequences at the C4d regions for the C4B 1 gene in linkage to C4A 6, and for the iso-expressed haplotype C4A 3 C4A 2 (C4B Q0, Yu and Campbell, 1987). For the first case, N 1157 (Rg2 / Ch -6) is expected as the C4B 1 protein expressed from the C4A 6 C4B 1 haplotype is Ch -3 and Ch -6. The C4 proteins from the C4A 3 C4A 2 (C4B Q0) haplotype have been found to be Rg 1,-2; Ch 5, 6 (Giles, 1987). If Rg2 and Ch6 are alternative, and because negative traits of the Rg and Ch

Fig. 5.1 A structural model for the location of the Rg and the Ch antigenic determinants and the correlation with the human complement C4A/C4B isotypes

The C4d fragment is from position 938 to 1317. The isotypic residues of C4A and C4B are marked by crosses. The eight amino acid changes at regions I, II, III and IV constitute Ch 5, Ch 4, Ch 6 and Ch 1, which are probably continuous epitopes. In addition they may form two discontinuous epitopes, Ch 2 and Ch 3. The location of the Rg determinants are derived from serological observation that on a C4 molecule, Rg 1 and Rg 2 are mutually exclusive to Ch 1 and Ch 6, respectively; and from the structural and serological data of C4B 5 and C4A 1. Deduced/assumed locations are enclosed in brackets. Solid triangle shows the position of intron in genomic-derived sequences. Open square shows the location of the thiolester residues.



determinants are expressed in a recessive manner, S 1157 (Ch6 / Rg-2) in the region III are therefore expected to be present on both of the C4A 2 and C4A 3 molecules, and G 1054 (Ch5) in the region I may be present on either or both of these iso-expressed C4A molecules. Alternatively the model can be tested by using synthetic peptides with amino acid sequences corresponding to the proposed antigenic sites for competition binding to the alloantibodies. On the other hand, the model is in accord with all of the serological data and no exception has been found so far (Giles, C.M. personal communication).

5.4 Prospective and implications

The elucidation of the C4A/C4B isotypic residues significantly advanced our understanding on the structure and function between two highly homologous proteins which exhibit diversified reactivities. It suggests the presence of a region in the C4 molecule that may modulate the covalent binding reaction of the thiolester carbonyl group (Yu et al., 1986). The determination of the molecular basis of Rg / Ch antigenic determinants would provide insights to the allo-antigenicity of human proteins. It is interesting to observe that only a single amino acid difference on a 1745-residue human serum protein can stimulate an immune response after blood transfusion. For instance, a transfused patient with C4A 4 C4B 2 haplotype (Rg 1, 2; Ch 1, -2, 3, 4, -5, 6) produced antiserum specific for Ch2 and Ch5 (Giles et al., 1987). In this aspect the Rg/Ch antigenic determinants of C4 are remarkably similar to the genetic markers in the

constant regions of the immunoglobulin molecules, Gm (reviewed in Nisonoff et al., 1975). The presence of multiple antigenic determinants on the same molecule, and some identical determinants on different isotypes, and the polyspecificities of the corresponding alloantisera are all common features shared by the two groups of serum proteins. In comparison, study of the antigenicities of C4A and C4B has the advantage of a manageable complexity.

The high immunogenicity of a protein region would suggest that this region is exposed or readily accessible for antibody-antigen interaction (Novotny et al., 1987). This infers that the C4d region may be located on the surface of the C4 molecule. A characteristic antigenic site of a protein probably spans about 6 amino acid residues (Kabat, 1970; Schechter, 1971), and they can be continuous (sequential) or discontinuous (conformational) in nature (Benjamin et al., 1984; Van Regenmortel, 1987). Therefore, the proposed locations for Rg/Ch determinants stand for stretch(es) of amino acids at which the polymorphic residue(s) are located. The proposed conformational epitopes Ch2 and Ch3 would infer that G 1054 and LSPVIH 1101-6 are reasonably proximal to each other, and so is S 1157 and ADLR 1188-91. As already noted, the C4A isotypic residues appear non-immunogenic. This might be the result of the unique, and probably very rigid structure formed by a cysteine and two proline residues. It was suggested that the C4A-isotypic residues cause the retardation of the alpha chain mobility (therefore, larger apparent mol. wt.) in modified SDS-PAGE (Roos et al., 1982; Chan and Atkinson, 1985). It is unclear why a parallel conformational epitope has not been formed between

N 1157 and VDLL 1188-91, although a novel determinant that requires Ch6 (i.e. S 1157) and Rg1 (i.e. VDLL 1188-91) in cis configuration has been described (Giles and Jones, 1987). The presence of this 'mixed' epitope offers considerable support, at least conceptually, to the proposed model.

A monoclonal antibody which can distinguish between Ch1 and Rg1 has been described (Dodds et al., 1986; Giles and Ford, 1986). These antibodies have proved useful in the separation of many C4A and C4B allotypes (Dodds et al., 1985, 1986).

Using a C4d-specific probe for Southern blot analysis, the presence of DNA sequence determining the expression of Rg1 / Ch1 epitopes and the C4A/C4B isotypic residues can be detected, respectively, by the EcoO 109 and Nla IV restriction fragment length polymorphisms (RFLPs) (Chapter 6; Yu and Campbell, 1987). The application of these definitive RFLPs enabled the studies of C4 genetics at DNA level that has proved useful in elucidation of the molecular basis of the C4 null alleles. It is also possible to distinguish between the Rg2 and Ch6 epitope at the genomic level by the Alu I RFLP at the corresponding position, if a recently available FCM agarose is used to separate the low mol. wt. DNA. The DNA sequence determining the Ch5 epitope cannot be recognized by a known restriction enzyme. However, its presence can be characterised by a specific oligonucleotide probe or by direct nucleotide sequencing. Application of these recombinant DNA techniques would be helpful in the typing of C4 that is important in the studies of the HLA genetics.

5.5 Summary

Serological typing by haemagglutination inhibition test has defined two Rg and six Ch antigenic determinants which are located in the C4d fragments of the fourth complement components C4A and C4B. Analysis of the sequence data at C4d among many cloned C4 genes from both isotypes (of which some have been serologically and biochemically characterised) has enabled the construction of a model for the exact locations and interrelationships of the C4A and C4B isotypic residues, and the Rg and Ch determinants. It is suggested that the eight amino acid substitutions located at four regions of the C4d domain in the C4B constitutes the Ch 1, 4, 5, and 6 determinants which are probably continuous (or sequential) epitopes; and that these four regions also constitute two discontinuous (or conformational) epitopes, Ch2 and Ch3. The locations of the Rg1 and Rg2 epitopes were deduced from the serological observation that on the C4 molecule, Rg1 and Rg2 appear mutually exclusive to Ch1 and Ch6, respectively. The model gives explanation to the complex relationships among the C4A and C4B allotypes and the Rg and Ch antigens. No violation to the model has yet been found.

CHAPTER SIX

DEFINITIVE RFLPS TO DISTINGUISH BETWEEN THE HUMAN COMPLEMENT C4A/C4B ISOTYPES AND THE MAJOR RODGERS/CHIDO DETERMINANTS: APPLICATION TO THE STUDY OF C4 NULL ALLELES

6.1 Introduction

The structural basis for the difference between the two isotypes of human C4, C4A and C4B, has been clarified recently. Four isotypic residues have been found between these two highly homologous but polymorphic molecules, each consists of 1745 amino acids (mol. wt. ~200,000). At position 1101-6 of the pro-C4 molecule, C4A can be defined by the presence of the sequence PCPVLD, and C4B by the presence of the sequence LSPVIH, respectively (Yu et al., 1986; Belt et al., 1984; 1985). These four isotypic residues are probably the origin of the different mobilities of the α chains of C4A and C4B on SDS-PAGE a(C4A- α 96,000 Mr; C4B- α 94,000 Mr; Roos et al., 1982), and is the basis for their differential chemical reactivities, for example, haemolytic activity and covalent binding affinities to peptide and carbohydrate antigens (see Porter, 1985, for a review). The four isotypic amino acid differences are caused by 5-nucleotide changes within seventeen nucleotides of DNA that appears as a entire block alteration. No other combination of change at this region has been found (chapter 4, Yu et al., 1986).

The Rodgers (Rg) and the Chido (Ch) blood group antigens were shown to correlate with the C4d fragments of C4A and C4B, respectively (O'Neill et al., 1978a; Tilley et al., 1978; Lundwall et al., 1982). However, reversed associations of Rg

with C4B (e.g. C4B 5) and Ch with C4A (e.g. C4A 1) have been found (Rittner et al., 1984; Roos et al., 1984). Two Rg and six Ch antigenic determinants have been defined and the major determinants of are Rg1 and Ch1 (Giles et al., 1984; Giles, 1987). Molecular cloning and DNA sequencing of the C4d polymorphic regions of the genes encoding C4A 1 and C4B 5 have enabled the location of the Rg1 and Ch1 determinants to position 1188-91 of the C4 molecules. The expression of the Rg1 epitope is determined by the sequence VDLL, and Ch1 by the sequence ADLR at that location. These variations in amino acid sequence are the results of 4 nucleotide changes within sixteen nucleotides of DNA, which, similar to those for the isotypic residues is the result of an entire block alteration.

An unusual phenomenon in the genetic of human C4 is the exceptionally high frequency of silent or null alleles, C4 Q0 (quantity zero), at either locus (frequency: C4A Q0 ~5-15%, C4B Q0 ~10-20%, Schendel et al., 1984; Tokunaga et al., 1985; Partansen and Koskimies, 1986). A null allele is defined as the absence of C4A or C4B protein in the serum (O'Neill et al., 1978b; Awdeh and Alper, 1980). Carroll and co-workers (Carroll et al., 1985a,b; 1986; Schneider et al., 1986) have shown that a proportion (~60% in their recent study) of the C4 haplotypes carrying null alleles are due to deletion of the C4 genes, usually together with their flanking 21-hydroxylase genes. However, the basis for the non-deleted C4 Q0 alleles remain unsolved. In a recent study the Taq 1 polymorphisms at the 5' region have been used to provide genetic markers for various C4 genes, i.e. C4A, C4B-long, C4B-short, and the C4B-short gene in association with the complotype SC01 (Schneider et

al., 1986), although the region determining the isotypicity of C4 lies closer to the 3' end of the C4 gene. The variation of the C4 gene size (i.e. C4-long, C4-short) is due to the absence of a 6-7 kb intron in some of the C4 (or C4B) genes (Carroll et al., 1985a; Yu et al., 1986).

In the present study, definitive RFLPs representing the exact locations corresponding to the C4A/C4B isotypic changes and to the Rg1/Ch1 antigenic determinants have been designed. This newly developed method is used to examine the structural basis of C4 Q0 alleles, particularly the 'non-deleted' or 'unexpressed' null alleles. The Taq 1 polymorphisms at the 5' end of C4 have also been included in this study and it appears that the polymorphism represents the nature of the two C4 loci instead of C4A/C4B genes, although they are generally correlated. It has been shown that the C4B Q0 allele of the HLA haplotype B44 DR6 C4A 3; C4B Q0 is not a C4B gene by definition, but very probably another C4A 3 allele at the second C4 locus of the same chromosome. It has been found that individuals carrying null alleles and some rare alleles of C4 are more susceptible to many HLA-associated disease (for a review see Hauptmann et al., 1986) such as systemic lupus erythematosus (Fielder et al., 1983; Reveille et al., 1985), Graves disease, and insulin depending diabetes mellitus (Partanen et al., 1986; Skanes et al., 1985b). Whether these diseases are caused by different efficiencies of the polymorphic C4 allotypes in the complement activation pathway (Porter, 1983), by deficient C4A or C4B protein itself, or by some other malfunctional (or susceptibility) genes in linkage disequilibrium with C4 null or rare alleles is yet to be

determined. The definitive RFLPs for C4A/C4B and Rg1/Ch1 shown here, together with the application of the locus specific 5' Taq I polymorphism of C4, may help to clarify the situation.

6.2 Result

6.2.1 C4A/C4B isotypicity - Nla IV polymorphism

C4A and C4B can be defined by four isotypic amino acid differences at position 1101-6, i.e. PCPVLD for C4A and LSPVIH for C4B. These class specific residues 5-nucleotide changes within stretch of seventeen nucleotides that is due to an entire block alteration. The C to T transition that attribute to the Pro/Leu difference between C4A and C4B generates an additional Nla IV site in the C4A, and extra Mnl I site in C4B, respectively. If the C4d genomic probe, P_B, is used for Southern blot analysis, C4A can be characterised by the presence of the 276 bp and 191 bp Nla IV or the 146 bp Mnl I restriction fragments, while C4B can be characterised by the presence of the 467 bp Nla IV or the 76 bp and 70 bp Mnl I fragments (Fig. 6.1). The very small Mnl I fragments are technically difficult to be detected by conventional method. Definitive identification of C4A and C4B by the Nla IV polymorphism is illustrated in Fig 6.1c. as shown in lanes 3 and 5, Mr. D (C4A 3; C4B Q0/C4A 3; C4B Q0) and his son A.D (C4A 1; C4B Q0 / C4A 3; C4B Q0) possess only the C4A-specific 276 bp and 191 Nla IV fragments, while JM (C4A 4; C4B 5/ C4A 3; C4B Q0, lane 2), Mrs D (C4A 3; C4B 3/C4A 3; C4B Q0, lane 4), and the hepatoma cell line Hep G2 (lane 1) possess the C4B-specific 467 bp Nla IV fragment, in addition to the C4A-specific 276 bp

Table 6.1 HLA haplotypes of cell-lines and individuals used in this study

Cell-lines	HLA		C2 FB		C4	Rg1/Ch1 phenotype		References
i) Ice 6	A2	B7	DR2	C	S	A3 BQ0		Carroll et al., 1985a; b
ii) Mann	A29 B44	DR7	F	AQ0 B1				de Kretser et al., 1982
Individuals ⁺								
1. Mr. D	a	A28 B51 C-	DR5 C	S	A3 BQ0	Rg1, Ch-1		Bentley et al., 1985
	b	A2 B44 C5	DR6 C	F	A3 BQ0(?)			
2. Mrs. D	c	A2 B62 C3	DR4 C	S	A3 B3	Rg1, Ch1		Bentley et al., 1985
	d	A3 B47 (C6)	DR7 C	F	A1 BQ0			
3. AD	b	A2 B44 C5	DR6 C	F	A3 BQ0(?)	Rg1, Ch1		Bentley et al., 1985; Yu et al., 1986
	d	A3 B47 (C6)	DR7 C	F	A1 BQ0			
4. JM	A3 B35 A24 B18	DR2(?) DR2(?)	Rg1, Ch-1		A3 BQ0 A4 B5			Yu et al., 1986

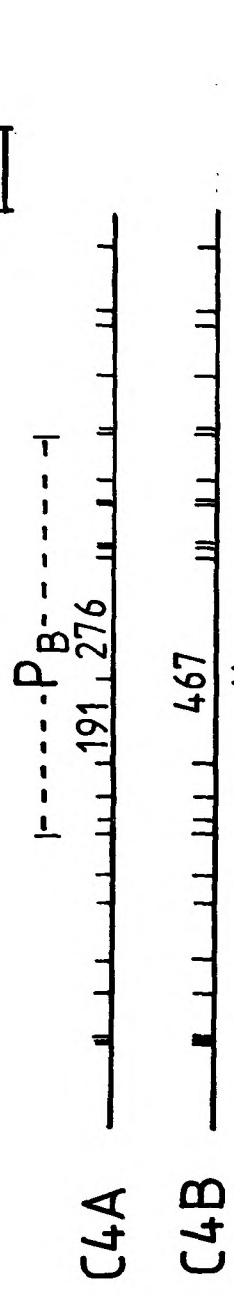
⁺ individuals 1-3 are members of family Q

Legend to Figure 6.1

Nla IV polymorphism for C4A/C4B isotypic determinants

- (a) Comparison of Nla IV restriction maps between C4A and C4B at the C4d region. Restriction sites were based on the genomic DNA sequence of a C4A 3 gene and a C4B 3 gene (Yu et al., 1986; Yu, C.Y. unpublished). Size of the polymorphic fragments are marked. P_B = genomic C4d specific probe used for hybridisation.
- (b) DNA and derived amino acid sequences at the C4 isotypic region. Isotypic amino acid residues are circled; the 5-nucleotide block alternations in the seventeen nucleotide block are marked by asterisks. Recognition sites of Nla IV and Mnl I are boxed. aa no. = amino acid number according to the pro-C4 molecule.
- (c) Genomic Southern blot analysis of the Nla IV polymorphism. C4 haplotypes of the DNA used are tabulated. A C4A gene can be defined by the 276bp and 191bp Nla IV fragments, and a C4B gene by a single 476bp Nla IV fragment, when P_B is used for hybridisation. ND = not determined.

200bp



a)

b)

aa no. 1099 1108

Q D (P) (C) P V (L) (D) R S

C4A CAGGACCCCTGTCCAGTGTTAGACAGGAGC

NlaIV * * * *

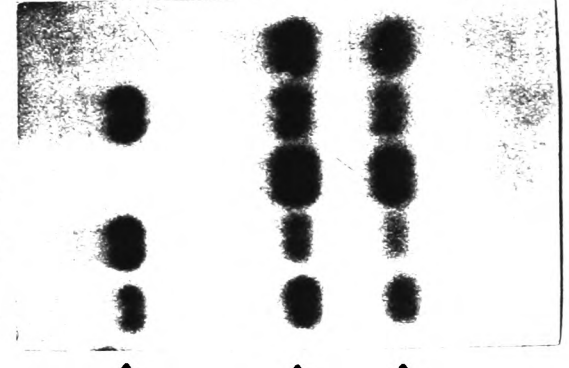
Mnl1

C4B CAGGACCCTCTCTCTCCAGTGATACATAGGAGC

Q D (L) (S) P V (I) (H) R S

c)

bp		Lanes	C4 haplotypes
C4B-467 →	1	Hep G2	ND
	2	JM	A4 B5/A3 BQ0
C4A — { 276 → 191 →	3	Mr D	A3 BQ0/A3 BQ0
	4	Mrs D	A1 BQ0/A3 B3
	5	A D	A1 BQ0/A3 BQ0



abnd 191 bp Nla IV pattern. It has also been found that the HLA homozygous cell lines, Ice 6 (C4A 3; C4B Q0), and Mann (C4A Q0; C4B 1) exhibit the C4A-specific and the C4B-specific Nla IV pattern, respectively. The HLA haplotypes of the individuals and cell lines described above are listed in Table 6.1.

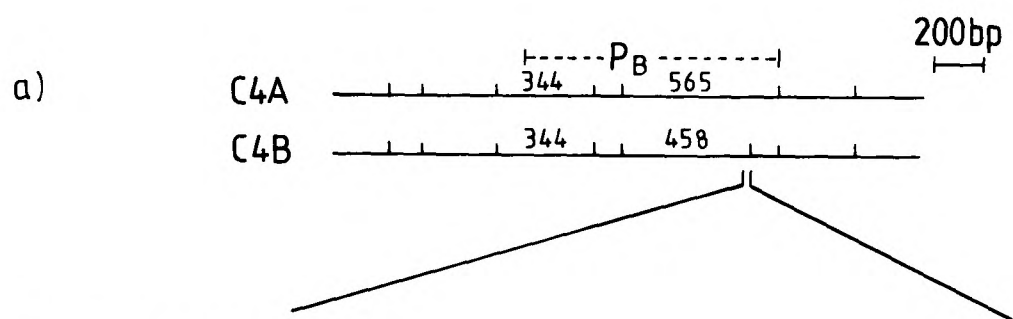
6.2.2 Rg1 and the Ch1 antigenicity - Eco0 109 polymorphism

Structural data from the variants C4A 1 and C4B 5 and many other allotypes suggested that the major C4 antigenic determinants Rg1 and Ch1 are determined by the sequence VDLL or ADLR, respectively, at position 1188-91 (Yu et al., 1986). The two dissimilar amino acids are caused by four nucleotides, which again appears as an entire block alteration, and no exception to this has been found (Fig 6.2a, b). The first of these nucleotide changes, a G to C transversion at codon 1186, generates an extra Eco0 109 restriction site that can be used as a diagnostic marker for the three other succeeding mutations involved in the expression of the Rg1 or Ch1 antigenic determinants. When the same C4d specific genomic probe, P_B, is used for Southern blot analysis, the C4 gene expressing Rg1 can be characterised by a 565 bp Eco0 109 fragment, while a Ch1 expressing C4 gene can be characterised by a 458 bp Eco0 109 fragment. As shown in Fig. 6.2, only the Rg1 specific 565 bp Eco0 109 fragment is seen in the individual JM (serologically typed Rg1, Ch1, lane 4). On the other hand, both the Rg1 specific 565 bp Eco0 109 and the Ch1 specific 458 bp Eco0 109 fragments are found in Mrs. D (lane 6), AD (lane 1) and Mann (lane 2) were tested, the Rg1-specific pattern is found in Ice 6 and the Ch1-specific pattern is found in Mann, suggesting that

Legend to Figure 6.2

Eco0 109 polymorphism for Rg1/Ch1 antigenic determinants

- (a) Comparison of Eco0 109 restriction map between a Rg1-expressing C4A and a Ch1-expressing C4B gene at the C4d region. Size of the Eco0 109 fragments picked up by P_B are marked.
- (b) DNA and derived amino acid sequences at the Rg1/Ch1 determining region. Amino acid residues for Rg1/Ch1 antigenic determinants are circled. The 4-nucleotide alterations in the sixteen nucleotide block responsible for Rg1/Ch1 expression are marked by asterisks. The recognition site of Eco0 109 is boxed.
- (c) Genomic Southern blot analysis of the Eco0 109 polymorphism. Rg1/Ch1 phenotypes of the cell lines/individuals from whom DNA was digested are tabulated. Derived phenotypes are enclosed by brackets. A Rg1-expressing C4 gene can be defined by a 565 bp fragment, and a Ch1-expressing C4 gene by a 458bp fragment, when P_B is used for hybridisation. Abbreviation and symbols used are the same as Fig.6-1.



b)

aa no. 1185 1193

Rg 1 K A P (V) D L (L) G V

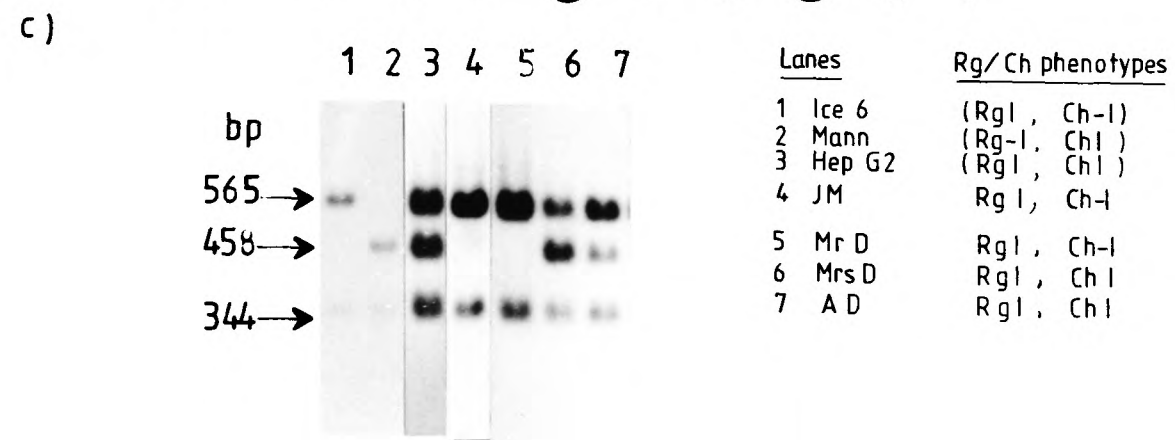
AAAGGCGCCGTGTGGACCTGCTCGGGTGTT

NarI * * *

Eco0109 *

AAGGCCCTGCGGACCTGCGGGGTGTT

Ch 1 K A P (A) D L (R) G V



these two cell-lines possess the Rg1- and Ch1- expressing C4 genes, respectively.

Besides being recognised by the restriction enzyme EcoO 109, the same transversional mutation at codon 1186 can be distinguished by Nar I (Yu et al., 1986) and Ban I. However, genomic DNA at the region is resistant to Nar I digestion, while the polymorphic restriction fragments generated by a Ban I digest cannot be as well-separated on agarose gels as the EcoO 109 digest.

6.2.3 Applications: studies on the C4 null alleles

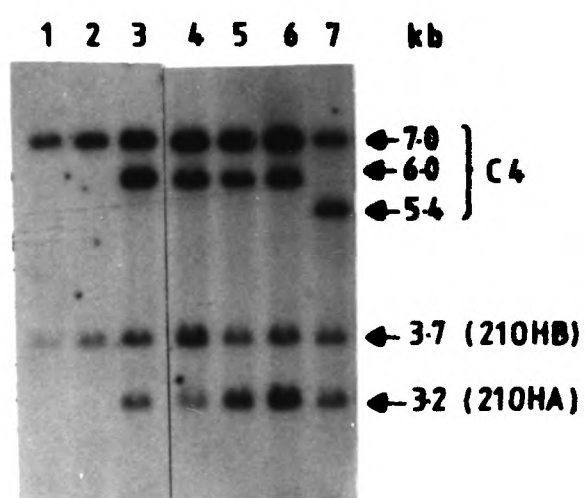
Schneider and coworkers (1986) have analysed the Taq I polymorphism of 126 MHC haplotypes of the human C4, 21-hydroxylase gene complex. For the case of C4, it was suggested that the 7.0 kb, 6.0 kb, 5.4 kb and 6.4 kb Taq I fragments detected by the 5' C4 cDNA probe (i.e. P_A) correspond to the long C4A, long C4B, short C4B, and short C4B genes with a deleted C4A Q0 allele of the SC01 complotype, respectively. However, the correlation of the 7.0 kb and the 6.0 kb Taq I fragments with C4A (long) and C4B (long) genes, respectively, seems not absolute, as revealed by the combined analysis of the Taq I polymorphism, and the Nla IV and EcoO 109 definitive RFLP studies already described on the two cell-lines and three members of a family (Q).

As shown in Fig. 6.3 (lane 2), only the 'C4A-like' 7.0 kb Taq I fragment can be detected from the cell line Mann (C4A Q0, C4B 1, homozygous). The Nla IV and EcoO 109 restriction patterns have already suggested that this cell line possesses only the C4B

Legend to Figure 6.3

Taq I polymorphism of the C4 gene loci and 21-hydroxylase genes

The 5' specific C4 cDNA probe, P_A, and the 3' specific 21-hydroxylase genomic probe, were used to analyse C4 gene loci and 21-hydroxylase genes A and B, respectively. C4 gene locus I is normally flanked by a 7.0kb 5' Taq I fragment and a 3.2kb 21-hydroxylase A 3' Taq I fragment, while the C4 gene locus II by the corresponding 6.0/5.4kb 5' fragment and 3.7kb 21-hydroxylase B 3' fragment. Notice (1) the C4B 1 gene of Mann (lane 2) is flanked by the locus 1 specific 7.0kb Taq I fragment; and (2) Mr. D (lane 4) and AD (lane 6), both with silent C4B Q0 alleles, possess the C4 locus II specific 6.0kb Taq I fragment, with band intensity about one half of the locus I specific 7.0kb Taq I fragment. See the neighbouring table and text for further description. 210H, 21-hydroxylase.



<u>Lanes</u>	<u>C4 haplotypes</u>	<u>Remarks</u>
1) Ice 6	A3BQ0, homozygous	210HA-C4B deleted
2) Mann	AQOB1, homozygous	3'-C4A-210HA-5'C4B deleted
3) Hep G2	ND	equimolecular C4A-210HA-C4B-210HB
4) Mr. D	A3BQ0/A3BQ0	210HA-C4B deleted on one chromosome, intact C4 locus II present on other.
5) Mrs. D	A1BQ0/A3B3	intact C4 locus II is present on one chromosome.
6) AD	A1BQ0/A3BQ0	intact C4 locus II is present on one chromosome.
7) JM	A4B5/A3BQ0	B5 is encoded by a short (i.e. 16 kb) C4 locus.

and Ch1-expressing gene. Further analysis using a human 21-hydroxylase genomic probe reveals that only the 21-hydroxylase B gene is present in the genome (as shown by the 3.7 kb Taq I fragment). Thus the C4B 1 gene of this particular haplotype may be a recombinant formed by the fusion of the 5' end of the first C4 locus to the 3' end of the second C4 locus (Carroll et al., 1985c). This anomolous association of a 7.0 kb Taq I fragment with C4B was also noted in one case by Schneider et al., (1986).

Phenotyping of the members of the family Q shows that no C4B protein could be detected in AD and his father, Mr. D, and therefore they were assigned homozygous C4B Q0 (Table 6.1). Southern blot analysis of the genomic DNA from these individuals showed that both the C4A-related 7.0 kb Taq I fragment and the C4B-related 6.0 kb Taq I fragments can be detected, but all with a ratio of 2:1 in band intensity. This would suggest that one of the C4B Q0 alleles at the C4B locus is deleted and the other is present but 'unexpressed'.

Definitive RFLP analyses for C4A/C4B isotypes and for Rg1/Ch1 expressing C4 genes on these individuals were carried out. For Mr. D., neither C4B nor a Ch1-expressing gene can be detected (Fig. 6.1, lane 3; Fig. 6.2, lane 5). For AD, no C4B-specific 467 bp Nla IV fragment was observed (Fig. 6.1, lane 3), and the ratio of the band intensity of the Rg1- to Ch1-specific EcoO 109 fragment is about 2:1 (Fig. 6.2, lane 7). Since the Ch1-specific EcoO 109 fragment corresponds to the presence of the C4A 1 gene in his genome, these altogether infer that the putative 'unexpressed' C4B Q0 allele of the HLA haplotype B44 DR6 is non-existing, although the second C4 locus is present, as

judged by the presence of the 7.0 kb and 6.0 kb Taq I fragment. This second C4 locus may express the C4A 3 allotype, identical to that of the first C4 locus on the same chromosome. Molecular cloning, restriction analyses and DNA sequencing of the C4 genes isolated from the AD genomic library is in favour of this suggestion (Yu et al., 1986; C.Y. Yu and others, in preparation).

The genuine non-expressed C4 null allele, which should possess the C4A or C4B specific Nla IV band pattern, but with the corresponding protein undetected, is yet to be established.

6.3 Discussion

6.3.1 RFLPs of human C4

RFLPs representing the exact locations of the nucleotide changes responsible for the human C4A/C4B isotypic differences and the Rg1/Ch1 antigenic determinants are described here. A C4d specific, 927 bp BamH I genomic probe (P_B) which includes sequence encoding the isotypic and Rg/Ch antigenic determinants, has been used for Southern blot analysis. A C4A gene is recognised by the 267 bp and the 191 bp Nla IV fragments and a C4B gene by a single 467 bp Nla IV fragment. In addition, the Rg1-expressing C4 gene will give a 565 bp EcoO 109 fragment, while a Ch1-expressing C4 gene gives a 458 bp EcoO 109 fragment. All these polymorphic restriction fragments can be unambiguously detected by means of electrophoresis on 1.5% agarose gel and the application of nylon membranes for the Southern blotting process. I anticipate this technique to be widely applied for the study of human C4 polymorphism in parallel with the well-established

methods at the protein level such as the α chain typing (Roos et al., 1982), serological typing for Rg/Ch antigens (Giles, 1985b, 1987), and the recently improved phenotyping of C4 using carboxypeptidase B treated plasma and immunofixation agarose gel electrophoresis (Sim and Cross, 1986; Mauff et al., 1983a; Awdeh and Alper, 1980). The method described here is efficient and convenient. It may give further definite information on whether the C4A/C4B gene of interest is present in the genome in spite of its nature (i.e. no matter whether it is functional or non-functional). It may also enable a rough estimation of the ration of different C4 genes present, although this has to be interpreted carefully. This is because the polymorphic restriction fragments are smaller than the probe (P_B) and the larger restriction fragments may bind to the probe more effectively than the smaller ones.

A summary of the diagnostic RFLPs for the human C4 genes is given in Table 6.2. As illustrated in this study, a combined genomic Southern analysis of the Taq I polymorphism (using the 5' cDNA probe P_A), and the Nla IV and EcoO 109 polymorphic pattern can give answers to the nature and characteristics of the two C4 gene loci such as the presence of the 6-7 kb intron, C4A or C4B, Rg1- or Ch1- expressing, and their relative gene dosages.

Table 6.2 A summary of diagnostic RFLPs for human C4 genes

	Restriction Enzyme	Probe	Band Size	Remarks	References
1) Long/short C4 genes (i.e. presence/absence a 6-7 kb intron)	Bam HI	5' cDNA	4.8 kb 3.5 kb	Long gene ⁺ Short gene [†]	Carroll et al., 1985a; b; Yu et al., 1986
	Kpn I	C4d genomic (or pALU-7)	7.5 kb 8.5 kb	Long gene Short gene	
2) Nature of Locus and probably presence of the 6-7 kb intron	Kpn I	5' cDNA	12 kb 3.5 kb	locus I, or I/II recombinant locus II	Carroll et al., 1985a
	Taq I	5' cNDA	7.0 kb 6.0 kb 5.4 kb 6.4 kb	locus I, long I/II recombinant locus II, long locus II, short locus II, short of SC01 complotype (possibly I/II recombinant)	
3) C4A or C4B gene	Nla IV	C4d genomic	(276+191)bp 467 bp	C4A gene C4B gene	Yu et al., 1986; this work
4) Rg1/Ch1 determinants	Eco0 109	C4d genomic	565 bp 458 bp	expresses Rg1 expresses Ch1	Yu et al.. 1986; this work

⁺ long gene, 22 kb in size, with a 6-7 kb intron

[†] short gene, 16 kb in size, without a 6-7 kb intron

6.3.2 C4 Q0 alleles

Genomic Southern analysis of the family Q for the Taq I, Nla IV and EcoO 109 RFLPs, together with the molecular cloning and DNA sequencing of the C4 genes from AD, suggest that the 'C4B-like' gene of the HLA haplotype B44 DR6 is more likely to be a gene encoding the identical isotype and allotype as its neighbouring locus, C4A 3, than the putative, silent, C4B Q0. In this case, the C4 haplotype is probably C4A 3; C4A 3 rather than C4A 3; C4B Q0. This is the first direct structural indication that the two C4 gene loci can encode identical allotypes, although the similar case for the same C4 isotype (e.g. HLA B35 DR1 C4A 3 C4A 2) has been well-documented (Fielder et al., 1983; Rittner et al., 1984; Schendel et al., 1984; Uring-Lambert et al., 1984).

Thus the C4 Q0 phenotype seems to be a heterogeneous mixture which consists of three distinct components: (1) deletion of the corresponding C4 gene locus which usually occur together with a 21-hydroxylase gene (Carroll et al., 1985a, b; 1986); (2) non-expression of a defective C4 mutant gene, which is yet to be established; and (3) expression of identical allotypes at the two C4 gene loci, with the HLA haplotype B44 DR6 C4A 3; C4B Q0 probably being an example. The third one would give an explanation for the exceptionally high frequency of the C4 Q0 allele in the population. Obviously a portion of the null alleles were assigned because the two C4 gene loci are expressing the phenotypically indistinguishable (identical) allotypes rather than the deleterious mutations or deletion of the C4 genes. This phenomenon is redefined as homo-expression so as to distinguish

it from iso-expression that involves the expression of C4 genes of the same isotype but different allotype (for example, C4A 3; C4A 2; C4B Q0) on one chromosome. The iso-expression (and homo-expression phenomenon was previously described as homo-duplication. Schneider and coworkers (1986) estimated that for the C4 null alleles, a ratio of 2 to 1 is found for deleted C4 genes to the 'unexpressed' C4 gene. The latter seems to be the sum of the malfunctional C4 genes and the homo-expressed C4 genes inferred in this study. The frequency of homo-expressing C4 genes in the C4 Q0 category remains to be determined. Knowledge of this may be helpful in elucidating the correlation of C4 Q0 alleles in the many HLA-related diseases already described.

6.3.3 The two loci of human C4

It was shown that there are generally two human C4 gene loci (O'Neill et al., 1978b) and that these loci encoded C4A and C4B respectively. (O'Neill et al., 1978a; Awdeh and Alper, 1980; Roos et al., 1982; Carroll et al., 1984a, b). In molecular terms, except for recombinant loci such as would be caused by deletion, the first locus (or locus I) can be characterised by a 7.0 kb Taq I fragment at the 5' end of the gene, and by the 21-hydroxylase A gene at the 3' flanking region; while the second locus (or locus II) can be characterised by a 6.0/5.4 kb Taq I 5' fragment and the 21-hydroxylase B gene at the 3' flanking region (Carroll et al., 1985c; White et al., 1985; Schneider et al., 1986; Table 6.2). Although generally the first locus encodes C4A and the second locus encodes C4B, exceptions may not be uncommon. Gene conversion or other recombinational events may have resulted

in the first locus expressing C4B or the second locus expressing C4A, as for the case of the reversed expression of the Rg/Ch antigenicities on the C4A 1 and C4B 5 allotypes (Yu et al., 1986). In this study, I have shown that the second C4 locus of the HLA haplotype B44 DR6 does not correspond to a C4B gene, but is probably encoding C4A 3 instead. Thus the two C4 loci may encode the same isotype, or more precisely the same allotype, of C4. Some of the 'homo-duplicated' C4 genes, e.g. C4A 3; C4A 2; C4B Q0 or C4A 5; C4A 2; C4B Q0 of the HLA haplotype A3 Cw4 B35 DR1 (Rittner et al., 1984) may well be due to expression of a C4A protein at the second locus and this suggests that the assignment of the C4B Q0 allele to these haplotypes can be confusing.

It remains to be established whether there is ever expression of the C4B isotype at the first locus, although it is clear that a I-II recombinant locus exists, such as that of the Mann cell-line, which maintains the locus I-specific 7.0 kb Taq I 5' fragment and expresses the C4B 1 allotype (the C4B 1 allele in this case is flanked by the locus II-specific 21-hydroxylase B gene at the 3' end).

6.4 Summary

Definitive restriction fragment length polymorphism (RFLPs) representing the exact locations responsible for isotypicity between the human complement component C4A and C4B, and their generally associated major Rodgers (Rg1) and Chido (Ch1) antigenic determinants, have been designed. By means of a C4d specific genomic probe for Southern blot analysis, a C4A gene can

be defined by the presence of the 276 bp and 191 bp Nla IV fragments, while C4B gene can be defined by a single 467 bp Nla IV fragment. In addition, a Rg1-expressing C4 gene can be represented by a 565 bp EcoO 109 fragment, and a Ch1-expressing C4 gene by a 458 bp EcoO 109 fragment, under the same conditions. All these polymorphic restriction fragments can be unambiguously and conveniently detected. In combination with the Taq I polymorphic patterns specific for the C4 gene loci and for the neighbouring 21-hydroxylase genes, the nature and structure of the tandem C4, 21-hydroxylase gene complex can be elucidated. In this study, it is inferred that the null allele of the HLA haplotype B44 DR6 C4A 3; C4B Q0 is not a C4B allele, but probably encodes another C4A 3 allotype at the second C4 locus.

CHAPTER SEVEN

GENERAL CONCLUSIONS

Molecular biology techniques have made significant contribution to the research of the complement system. During the past five years, cDNA clones for more than eighteen human complement proteins have been obtained and characterised (reviewed in Reid et al., 1987; chapter 1). Through cDNA sequencing, the complete or partial primary structures of the corresponding proteins were elucidated, which provides a solid foundation for further biochemical studies on their structure, function and genetics at the DNA level. Previous biochemical studies revealed many unusual features of the complement proteins, e.g. the circulation of the active forms of serine proteases in the plasma (factor D and factor I); the presence of the serine protease catalytic domain on a chain having approximately double the mol. wt. of other serine proteases (factor B and C2); the collagen-like protein (C1q); thiolester-containing proteins (C3 and C4); and the multiple-component complex enzymes (C3 and C5 convertases) (reviewed in Reid and Porter, 1981; Lachmann and Hugh-Jones, 1984). However, recent studies by recombinant DNA techniques and monoclonal antibody technology have started to unravel the structural relationships of several complement proteins with proteins of other systems. The most elegant results are the demonstration of a short consensus repeat of ~ 60 amino acids in more than eight complement regulatory proteins, and also in the interleukin 2 receptor, β_2 glycoprotein I and haptoglobin (reviewed in Reid et al., 1986); as well as the structural homologies among C8 α , C8 β , and C9, and their structural 'kinships' with the

epidermal growth factor and the low density lipoprotein receptor (chapter 1).

Structural homologies are most likely to have occurred as a result of divergent evolution from one or a few ancestral exons. The gene (exon-intron) structure has only been determined for two complement proteins, factor B (Campbell and Porter, 1984; Campbell et al., 1984), and the C1q B chain (Reid, 1985). This is partly due to a heavier work load as the gene sequences are generally 3-100 fold longer than their corresponding messenger RNAs (or cDNAs). C4 is now the third and the biggest (in terms of exon number) complement component analysed at the genomic level. The genomic DNA sequence of the C4A 3a gene has confirmed the published cDNA sequences (Belt et al., 1984), clarified some ambiguities of the cDNA sequences, elucidated the primary structure of the putative leader peptide, revealed the presence of more polymorphic sites and their locations in the C4 gene. It also localises the structural and functional aspects of the C4 protein to individual exons in many cases, such as the intra-cellular proteolytic site for the β - α and the α - γ chain junctions. the extracellular metalloprotease cleavage site near the carboxyl terminus of the α chain, the C1s and factor I-mediated cleavage sites, the glycosylations, sulphations, thiolester, and the C4A/C4B isotypic sites. When the exon-intron structures of the related proteins such as C3, C5 and the protease inhibitor α 2 macroglobulin are also determined, valuable information in their evolutionary relationships may be obtained. The fact that the anaphylatoxin C4a, and the thiolester site are encoded by phase 1-1 symmetrical exons have

helped to explain the absence of these structures in macroglobulin and C5, respectively. It could be due to the deletion or insertion of the corresponding symmetrical exons at the genomic level that would not change the reading frames of the coding sequences. Since exons may delineate supersecondary folding structures or specific functions of the protein, the information of the exon-intron structure could be more useful if the 3-dimensional structure of the C4 molecule is determined, and the protein functions of C4 are further characterised.

Knowledge of the precise locations of the polymorphic sites in the C4 gene, the availability of the peripheral blood samples from patients whose C4 proteins were biochemically (Dodds et al., 1986) and serologically (Giles, 1987) characterised, have led to the molecular cloning of many informative, rare and common C4 variants. Together these data have identified the C4A/C4B isotypic residues, and the Rodgers (Rg) and Chido (Ch) blood group antigenic determinants. The former offers an explanation for the differential thiolester binding reactivities of C4A and C4B. It infers the presence of a region on the C4 protein that would modulate the covalent binding affinities of C4, and probably in other thiolester-containing proteins. Further research using site directed mutagenesis in this region may firmly establish the role of these residues in the C4 molecule. The latter gives a satisfactory answer to the complex relationship among the expression of the two Rg and six Ch antigenic determinants, and further helps confirm the location of the amino acid residues determining the C4A and C4B isotypes.

The elucidation of the isotypic residues for C4A (PCPVLD 1101-6) and for C4B (LSPVIH 1101-6) have also enabled the designation of definitive RFLPs for these two genes due to differences in nucleotide sequences. Similarly, RFLPs corresponding to the major Rg/Ch antigenic determinants were identified. Together with other restriction digests which differentiated a variety of C4 and their 3' neighbouring 210Hase genes, the concept in the genetics of the C4 loci has been revised. In this work (Chapter 6), it has been shown that the two C4 loci of the HLA haplotype B44 DR6 C4A 3 C4B Q0 are probably encoding two identical isotypes (or more specifically, two identical allotype, C4A 3). In other words, the C4B Q0 allele of this haplotype is not a C4B gene by definition. Previously there was doubt over the number and nature of the gene loci of C4 on the haploid genome (Mauff et al., 1983a although O'Neill and coworkers (1978b) and Adweh and Alper (1980) rested the controversy by suggesting a C4A-C4B, two-locus model with a high frequency of null alleles at both loci (~10-20% at each locus). This hypothesis was mainly based on results from agarose typing gels (see chapter 1). However it is difficult to resolve the amount of the various allotypic proteins by band intensity as each allotype is represented by three protein bands. Recently Carroll and coworkers (Carroll et al., 1985b; Schneider et al., 1986) established that a proportion of the C4A/C4B null alleles (~60%) is due to the presence of only a single C4 gene locus (the other one may be deleted). Thus the null alleles of C4 are actually a heterogeneous mixture of (1) isoexpressed or homoexpressed C4 genes at both C4 loci, (2) presence of a single C4 locus only

on a haploid genome, or (3) the presence of a defective gene. As a result of this complexity, the two loci of C4, if present, are renamed locus I and locus II, instead of C4A and C4B. This will avoid confusion as either isotype of C4 may be expressed from either C4 locus.

In practical terms, the definitive RFLPs for C4A/C4B isotypes, for the Rg1/Ch1 determinants, and for the C4 locus I and locus II together may be useful in augmenting the typing of C4. This is because valuable information on the gene dosage and on the nature of the C4 gene loci can be determined through these RFLP studies. In parallel with the recently improved phenotyping method of C4 proteins on carboxypeptidase B treated plasma devised by Sim and Cross (1986), the genetics of human C4 can be investigated with greater accuracy and confidence. This may help to resolve the intriguing aspect of the apparent association of certain polymorphic forms of C4 with some HLA-linked diseases.

The mosaic DNA polymorphic pattern, the mixed expression of some Rg and Ch antigenic determinants on C4A and C4B, and the possible expression of C4A and C4B at either C4 loci altogether suggest that a gene conversion-like mechanism might have operated in mixing the many phenotypes of the C4 gene products from locus I and locus II. How this occurred is unknown, but further research on the mechanism for the variation of gene size at the second C4 locus might give some hints. The unusual variation in gene size is due to the presence of a bigger intron about 6-7 kb between exons 10 and 11. Preliminary results revealed the presence of direct long

terminal repeats flanking this big intron, suggesting that it could be some retroviral-like mobile element. It may be important to demonstrate whether it has biological activity. Studies on the gene structure of the 21OHase A and B genes and the molecular basis of congenital adrenal hyperplasia (a hereditary disease caused by 21OHase deficiency) have uncovered extensive polymorphism in these genes which are just located 3' to C4_L.^(Rodrigues et al., 1987) It would be worthwhile to determine whether the mechanism(s) attributing to the polymorphisms of the C4 and 21OHase genes are linked.

Thus although no obvious functional correlation and similarities among C4 and the MHC class I and class II products have been discovered, the sophistication of the C4 gene and protein structures and those of the classes I and II is approaching uniformity. Firstly, the genomic organisation of C4 is complex by varying the number of locus from one to three or even more (Yu, C.Y. and Dodds, A., unpublished observations) in the haploid genome. Secondly, most individuals contain more than one C4 locus and they express proteins with differential chemical reactivities, or they express proteins with variable Rg/Ch antigenic determinants. Thirdly, there is a collection of C4 alleles for each C4 locus. This complexity in gene number, diversity in gene products and polymorphism in allotypes are major characteristics of the MHC class I and II molecules (Steinmetz, 1984). C4 has further degrees of variation: heterogeneity in gene size at the same locus, and the production of many forms of post-translationally processed proteins in the plasma.

Recombinant DNA techniques have been applied to help elucidation of protein structure (and function) and they are extremely useful. However, this is only one of the approaches to study the molecule and the fundamental aspects on the biological roles of a protein cannot be neglected. Further advances on the knowledge of complement C4 (and other complement proteins) will require studies on the protein molecule itself, and on the functional roles it play in the body's defence system.

REFERENCES

- Allen, F.H.Jr. (1974) *Vox Sang.* 27, 382-384.
- Alper, C.A., Marcus, D., Raum, D., Petersen, B.H. and Spira, T.J. (1983) *J. Clin. Invest.* 72, 1526-1531.
- Alper, C.A., Awdeh, Z., Raum, D. and Yunis, E.J. (1986) *Biochem. Soc. Symp.* 51, 19-28.
- Alsenz, J., Schulz, T.F., Lambris, J.D., Sim, R.B. and Dierich, M.P. (1985) *Biochem. J.* 232, 841-850.
- Anson, D.S., Choo, K.H., Rees, D.J.G., Giannelli, F., Gould, K., Huddleston, J.A. and Brownlee, G.G. (1984) *EMBO J.* 3, 1053-1060.
- Arlaud, G.J. and Gagnon, J. (1983) *Biochemistry* 22, 1758-1764.
- Arlaud, G.J., Colomb, M.G. and Gagnon, J. (1987a) *Immunol. Today* 8, 106-111.
- Arlaud, G.J., Willis, A.C. and Gagnon, J. (1987b) *Biochem. J.* 241, 711-720.
- Awdeh, Z.L. and Alper, C.A. (1980) *Proc. Natl. Acad. Sci. USA* 77, 3576-3580.
- Ballard, L., Seya, T., Teckman, J., Lublin, D.M. and Atkinson, J.P. (1987) *J. Immunol.* 138, 3850-3855.
- Batchelor, J.R. and McMichael, A.J. (1987) *Br. Med. Bull.* 43, 156-183.
- Bell, G.I., Karam, J.H. and Rutter, W.J. (1981) *Proc. Natl. Acad. Sci. USA* 78, 5759-5763.
- Belt, K.T., Carroll, M.C. and Porter, R.R. (1984) *Cell* 36, 907-914.
- Belt, K.T., Yu, C.Y., Carroll, M.C. and Porter, R.R. (1985) *Immunogenetics* 21, 173-180.
- Benjamin, D.C., Berzofsky, J.A., East, I.J., Gurd, F.R.N., Hannum, C., Leach, S.J., Margoliash, E., Michael, J.G., Miller, A., Prager, E.M., Reichlin, M., Sercarz, E.E., Smith-Gill, S.J., Todd, P.E. and Wilson, A.C. (1984) *Ann. Rev. Immunol.* 2, 67-101.
- Bentley, D.R. and Porter, R.R. (1984) *Proc. Natl. Acad. Sci. USA* 81, 1212-1215.
- Bentley, D.R., Campbell, R.D. and Cross, S. J. (1985) *Immunogenetics* 22, 377-390.
- Bentley, D.R. (1986) *Biochem. J.* 239, 339-345.

- Bernards, R. and Flavell, R.A. (1980) *Nucleic Acid Res.* 8, 1521-1534.
- Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) *Proc. Natl. Acad. Sci. USA* 80, 3963-3965.
- Birnboim, H.C. and Doly, J. (1979) *Nucleic Acid Res.* 7, 1513-1523.
- Blake, C.C.F. (1978) *Nature* 273, 267.
- Blake, C.C.F. (1979) *Nature* 277, 598.
- Blake, C.C.F. (1983) *Nature* 305, 535-537.
- Blake, C.C.F. (1985) *Int. Rev. Cyt.* 93, 149-185.
- Borsos, T., Dourmashkin, R.R. and Humphrey, J.H. (1964) *Nature* 202, 251-252.
- Breathnach, R., Benoist, C., O'Hara, K., Gannon, F. and Chambon, P. (1978) *Proc. Natl. Acad. Sci. USA* 75, 4853-4857.
- Breathnach, R. and Chambon, P. (1981) *Ann. Rev. Biochem.* 50, 349-383.
- Campbell, R.D., Dodds, A.W. and Porter, R.R. (1980) *Biochem. J.* 189, 67-80.
- Campbell, R.D., Gagnon, J. and Porter, R.R. (1981) *Biochem. J.* 199, 359-370.
- Campbell, R.D. and Porter, R.R. (1983) *Proc. Natl. Acad. Sci. USA* 80, 4464-4468.
- Campbell, R.D., Bentley, D.R. and Morley, B.J. (1984) *Phil. Trans. R. Soc. Lond. B.* 306, 367-378.
- Campbell, R.D. and Bentley, D.R. (1985) *Immunol. Rev.* 87, 19-37.
- Campbell, R.D., Carroll, M.C. and Porter, R.R. (1986) *Adv. in Immunol.* 38, 203-244.
- Caras, I.W., Daritz, M.A., Rhee, L., Weddell, G., Martin, Jr, D.W. and Nussenzweig, V. (1987) *Nature*, 325, 545-549.
- Carroll, M.C. and Porter, R.R. (1983) *Proc. Natl. Acad. Sci. USA* 80, 264-267.
- Carroll, M.C., Campbell, R.D. Bentley, D.R. and Porter, R.R. (1984a) *Nature* 307, 237-241.
- Carroll, M.C., Belt, K.T., Palsdottir, A. and Porter, R.R. (1984b) *Phil. Trans. R. Soc. Lond.*, 306, 379-388.
- Carroll, M.C., Belt, K.T., Palsdottir, A. and Yu, C.Y. (1985a) *Immunol. Rev.* 87, 39-60.

- Carroll, M.C., Palsdottir, A., Belt, K.T. and Porter, R.R. (1985b) EMBO J. 4, 2547-2552.
- Carroll, M.C., Campbell, R.D. and Porter, R.R. (1985c) Proc. Natl. Acad. Sci. USA 82, 521-525.
- Carroll, M.C., Palsdottir, A., Belt, K.T. and Yu, C.Y. (1986) Biochem. Soc. Symp. 51, 29-36.
- Carroll, M.C., Katzman, P., Alicot, E.M., Koller, B.H., Geraghty, J.L. and Spies, T. (1987) Proc. Natl. Acad. Sci. USA (in press).
- Carter, P.E., Dunbar, B. and Fothergill, J.E. (1984) Phil. Trans. R. Soc. Lond. B. 306, 293-299.
- Casadaban, M.J. and Cohen, S.N. (1980) J. Mol. Biol. 138, 179-207.
- Catterall, C.F., Lyons, A., Sim, R.B., Day, A.J. and Harris, T.J.R. (1987) Biochem. J. 242, 849-856.
- Chakravarti, D.N., Campbell, R.D. and Gagnon, J. (1983) FEBS Letts, 154, 387-390.
- Chakravarti, D.N., Campbell, R.D. and Porter, R.R. (1987) Molec. Immunol. (in press).
- Chan, A.C. and Atkinson, J.P. (1983) J. Clin. Invest. 72, 1639-1649.
- Chan, A.C., Mitchell, K.R., Munns, T., Karp, D.R. and Atkinson, J.P. (1983) Proc. Natl. Acad. Sci. USA 80, 268-272.
- Chan, A.C., Karp, D.R., Shreffler, D.C. and Atkinson, J.P. (1984) Immunol. Today 5, 200-203.
- Chan, A.C. and Atkinson, J.P. (1985) J. Immunol. 134, 1790-1798.
- Chaplin, D.D., Woods, D.E., Whitehead, A.S., Goldberger, G., Colten, H.R. and Seidman, J.G. (1983) Proc. Natl. Acad. Sci. USA 80, 6947-6951.
- Christie, D.L., Gagnon, J. and Porter, R.R. (1980) Proc. Natl. Acad. Sci. USA 77, 4923-4927.
- Chung, L.P., Bentley, D.R. and Reid, K.B.M. (1985a) Biochem. J. 230, 133-141.
- Chung, L.P., Gagnon, J. and Reid, K.B.M. (1985b) Molec. Immunol. 22, 427-435.
- Clewell, D.B. and Helsinki, D.R. (1972) J. Bact. 110, 1135-1146.

- Colomb, M.G., Arlaud, G.J., and Villiers, C.L. (1984) Phil. Trans. R. Soc. Lond. B. 306, 283-292.
- Cooper, N.R. and Muller-Eberhard, H.J. (1970) J. Exp. Med. 132, 775-793.
- Cooper, N.R. (1975) J. Exp. Med. 141, 890-903.
- Cooper, N.R. (1985) Adv. Immunol. 37, 151-215.
- Cornacoff, J.B., Hebert, L.A., Smead, W.L., Van Aman, M.E., Birmingham, D.J. and Waxman, F.J. (1983) J. Clin. Invest. 71, 236-247.
- Craik, C.S., Sprang, S., Fletterick, R. and Rutter, W.J. (1982) Nature 299, 180-182.
- Craik, C.S., Rutter, W.J. and Fletterick, R. (1983) Science 220, 1125-1129.
- Crossley, L.G. (1981) Methods Enzymol. 80, 112-124.
- Dahlback, B., Smith, C.A. and Muller-Eberhard, H.J. (1983) Proc. Natl. Acad. Sci. USA 80, 3461-3465.
- Dawkins, R.L., Christiansen, F.T., Kay, P.H., Garlepp, M., McCluskey, J., Hollingsworth, P.N. and Zilko, P.J. (1983) Immunol. Rev. 70, 5-22.
- De Bruijn, M.H.L. and Fey, G.H. (1985) Proc. Natl. Acad. Sci. USA 82, 708-712.
- De Crombrughe, B. and Pastan, I. (1982) Trends Biochemi Sci. 7, 11-13..
- Deininger, P.L. (1983) Anal. Biochem. 129, 216-223.
- DeKretser, T.A., Crupmton, M.J., Bodmer, J.G. and Bodmer, W.F. (1982) Eur. J. Immunol., 12, 214-221.
- Denhardt, D.T. (1966) Biochem. Biophys. Res. Commun. 23, 641-646.
- D'Eustachio, P., Kristensen, T., Wetsel, R.A., Riblet, R., Taylor, B.A. and Tack, B.F. (1986) J. Immunol. 137, 3990-3995.
- Dierk, P., van Oeyen, A., Cochran, M.D., Dobkin, C., Reiser, J., and Weissmann, C. (1983) Cell 32 695-705.
- DiScipio, R.G., Smith, C.A., Muller-Eberhard, H.J. and Hugli, T.E. (1983) J. Biol. Chem. 258, 10629-10636.
- DiScipio, R.G., Gehring, M.R., Podack, E.R., Kan. C.C., Hugli, T.E. and Fey, G.H. (1984) Proc. Natl. Acad. Sci. USA 81, 7298-7302.
- Dodds, A.W., Law, S.K.A., and Porter, R.R. (1985) EMBO. J. 4,

2239-2244.

- Dodds, A.W., Law, S.K.A. and Porter, R.R. (1986) Immunolgenet. 24, 279-285.
- Doolittle, R.F. (1985) Trends Biochem. Sci. 10, 233-237.
- Duckworth, M.L., Gait, M.J., Goelet, P., Hong, G.F., Singh, M. and Titmas, R.C. (1981) Nucleic Acid Res. 9, 1691-1706.
- Dunham, I., Sargent, C. and Campbell, R.D. (1987) Proc. Natl. Acad. Sci. USA (in press).
- Dykman, T.R., Hatch, J.A. and Atkinson, J.P. (1984) J. Exp. Med. 159, 691-703.
- Fearon, D.T. (1979) Proc. Natl. Acad. Sci. USA 76, 5867-5871.
- Fearon, D.T. and Wong, W.W. (1983) Ann. Rev. Immunol. 1, 243-271.
- Fearon, D.T. (1984) Immunol. Today 5, 105-110.
- Feinberg, A.P. and Vogelstein, B. (1984) Anal. Biochem. 137, 266-267.
- Fielder, A.H.L., Walport, M.J., Batchelor, J.R., Rynes, R.I., Black, C.M., Dodi, I.A. and Hughes, G.R.V. (1983) Br. Med. J. 286, 425-428.
- Fishelson, Z. and Muller-Eberhard, H.J. (1982) J. Immunol. 129, 2603.
- Fishelson, Z., Pangburn, M.E. and Muller-Eberhard, H.J. (1984) J. Immunol. 132, 1430-1434.
- Frischauf, A-M., Lehrach, H., Poustka, A. and Murray, N.E. (1983) J. Mol. Biol. 70, 827-842.
- Fu, S.M., Kunkel, H.G., Brusman, H.P., Allen, F.H. and Fotino, M. (1974) J. Exp. Med. 148, 1044-1051.
- Fujita, T. and Nussenzweig, V. (1979) J. Exp. Med. 150, 267-276.
- Gagnon, J. (1984) Phil. Trans. R. Soc. Lond. B 306, 301-309.
- Garoff, H. and Ansorge, W. (1981) Anal. Biochem. 115, 450-457.
- Gergen, J.P., Stern, R.H., and Wensink, P.C. (1979) J. Immuno. 122, 146-151.
- Gigli, I., Von Zabern, I. and Porter, R.R. (1977) Biochem. J. 165, 439-446.
- Gigli, I. (1978) Nature 272, 836-837.
- Gigli, I., Fujita, T. and Nussenzweig V. (1979) Proc. Natl.

- Acad. Sci. USA 76, 6596-6600.
- Gigli, I., Sorvillo, J. and Halbwachs-Mecardli (1985) J. Immunol. 135, 440-444.
- Gil, A. and Proudfoot, N.J. (1987) Cell 49, 399-406.
- Gilbert, W. (1978) Nature 271, 501.
- Gilbert, W. (1985) Science 228, 823-824.
- Giles, C.M., Gedde-Dahl, T., Robson, E.B., Thorsby, E., Olaisen, B., Aenason, A., Kissmeyer-Neilsen, F., and Schreuder, I. (1976) Tissue Antigens 8, 143-149.
- Giles, C.M. (1980) In a Seminar on Antigens on Blood Cells and Body Fluids, American association of Blood Banks. pp 33-49.
- Giles, C.M. (1984) Vox Sang. 46, 149-156.
- Giles, C.M., Batchelor, J.R., Dodi, I.A., Fielder, A.H.L., Rittner, C., Mauff, G., Bender, K., Levene, C., Schreuder, G.M.T. and Wells, L.J. (1984) J. Immunogenet. 11, 305-317.
- Giles, C.M. (1985a) Vox Sang., 48, 160-166.
- Giles, C.M. (1985b) Vox Sang., 48, 167-173.
- Giles, C.M. and Ford, D.S. (1986) Transfusion 26, 370-374.
- Giles, C.M. (1987) Human Immunol. 18, 111-122.
- Giles, C.M. and Jones, J.W. (1987) Immunogenetics (in press).
- Giles, C.M., Hoffman, M., Moulds, M., Harris, M. and Dalmaso, A. (1987) Vox Sang. 52, 129-133.
- Go, M. (1981) Nature 291, 90-92.
- Go, M. (1983) Proc. Natl. Acad. Sci. USA 80, 1964-1968.
- Goldberger, G. and Colten, H.R. (1980) Nature 286, 514-516.
- Gorman, C.M., Moffat, L.F. and Howard, B. (1982) Mol. Cell Biol. 2, 1044-1051.
- Gorski, J.P., Hugli, T.E. and Muller-Eberhard, H.J. (1979) Proc. Natl. Acad. Sci. USA 76, 5299-5302.
- Gorski, J.P. and Muller-Eberhard, H.J. (1980) J. Immunol. 124, 1523 (abstr.).
- Gorski, J.P., Hugli, T.E. and Muller-Eberhard, H.J. (1981) J. Biol. Chem. 256, 2702-2711.
- Greer, J. (1985) Science 228, 1055-1060.

- Grosveld, F.G., Lund, T., Murray, E.J., Dahl, H.H.M. and Flavell, R.A. (1982) *Nucleic Acid Res.* 10, 6715-6732.
- Grubb, R. (1956) *Acta Pathol. Microbiol. Scand.* 39, 195-197.
- Haefliger, J-A., Tschopp, J., Nardelli, D., Wahli, W., Kocher, H-P., Tosi, M. and Stanley, K.K. (1987) *Biochemistry* 26, 3551-3555.
- Halbwachs, L., Leveille, M., Lesavre, Ph. Wattel, S. and Leibowitch, J. (1980) *J. Clin. Invest.* 65, 1249-1256.
- Hall, R.E. and Colten, H.R. (1977) *J. Immunol.* 118, 1903-1905.
- Harris, J.P., Tegoli, J., Swanson, J., Fisher, N., Gavin, J. and Noades, J. (1967) *Vox Sang.* 12, 140-142.
- Harrison, R.A., Thomas, M.L., and Tack, B.F. (1981) *Proc. Natl. Acad. Sci. USA* 78, 7388-7392.
- Hauptmann, G., Goetz, J., Uring-Lambert, B. and Gresshans, E. (1986) *Prog. Allergy* 39, 232-249.
- Hellman, U., Eggertsen, G., Lundwall, A., Engstrom, I. and Sjogquist, J. (1984) *FEBS Letts.*, 170, 254-258.
- Herrmann, R., Neugebauer, K., Pirkel, E., Zentgraf, H. and Schaller, H. (1980) *Molec. Genet.* 177, 231-242.
- Higashi, Y., Yoshioka, H., Yamane, M., Gotoh, O. and Fujii-Kuriyama, Y. (1986) *Proc. Natl. Acad. Sci. USA* 83, 2841-2845.
- Hing, S.N., Giles, C.M. Fielder, A.H.L. and Batchelor, J.R. (1986) *Immunogenetics* 23, 151-155.
- Hogness, D.S. and Simmons, J.R. (1964) *J. Mol. Biol.* 9, 411-438.
- Holers, V.M., Chaplin, D.D., Leykam, J.F., Gruner, B.A., Kumar, V. and Atkinson, J.P. (1987) *Proc. natl. Acad. Sci. USA* 84, 2459-2463.
- Holers, V.M., Cole, J.L., Lublin, D.M., seya, T. and Atkinson, J. (1985) *Immunol. Today*, 6, 188-192.
- Hortin, G., Chan, A.C., Fok, K.F., Strause, A.W. and Atkinson, J.P. (1986a) *J. Biol. Chem.* 261, 9065-9069.
- Hortin, G., Sims, H. and Strauss, A.W. (1986b) *J. Biol. Chem.* 261, 1786-1793.
- Howard, O.M.Z., Rao, A.G. and Sodetz, J.M. (1987) *Biochemistry* 26, 3565-3570.
- Huber, R., Scholze, H., Paques, E.P. and Deisenhofer, J. (1980) *Hoppe-Seyler's Z. Physiol. Chem.* 361, 1389-1399.

- Hugli, T.E. (1986) Complement 3, 111-127.
- Ichihara, C., Nakamura, T., Nagasawa, S. and Koyaya, J. (1986) Mol. Immunol. 23, 151-157.
- Isenman, D.E. and Kells, D.I.C. (1982) Biochemistry 21, 1109-1117.
- Isenman, D.E. and Young, J.R. (1984) J. Immunol. 132, 3019-3027.
- Isenman, D.E. and Young, J.R. (1986) J. Immunol. 136, 2542-2550.
- Janatova, J. and Tack, B. (1981) Biochemistry 20, 2394-2402.
- Janatova, J. (1986) Biochem. J. 233, 819-825.
- Jenne, D. and Stanley, K.K. (1985) EMBO J. 4, 3153-3157.
- Johnson, P.H. and Grossman, L.I. (1977) Biochemistry, 16, 4217-4224.
- Journet, A. and Tosi, M. (1986) Biochem. J. 240, 783-787.
- Kabat, E. (1970) Ann. N.Y. Acad. Sci. 169, 43-54.
- Kaiser, K. and Murray, N.E. (1985) in Glover, D.M. (ed.) DNA Cloning, vol. 1. IRL Press, Oxford, pp 1-47.
- Kan, C.C., Solomon, E., Belt, K.T., Chain, A.C., Hiorns, L.R. and Fey, G. (1985) Proc. Natl. Acad. Sci. USA 82, 2282-2286.
- Karn, J., Brenner, S., Barnett, L. and Cesareni, G. (1980) Proc. Natl. Acad. Sci. USA 77, 5172-5176.
- Karp, D.R. (1983) J. Biol. Chem. 258, 12745-12748.
- Kerr, M.A. (1980) Biochem. J. 189, 173-181.
- Kinoshita, T., Medof, M.E., Hong, K. and Nussenzweig, V. (1986) J. Exp. Med. 164, 1377-1388.
- Klickstein, L.B., Wong, W.W., Smith, J.A., Weis, J.H., Wilson, J.G. and Fearon, D.T. (1987) J. Exp. Med. 165, 1095-1112.
- Kristensen, T. and Tack, B.F. (1986) Proc. Natl. Acad. Sci. USA 83, 3963-3967.
- Kyte, J. and Doolittle, R.F. (1982) J. Mol. Biol. 157, 105-132.
- Lachmann, P.J. and Hobart, M.J. (1978) J. Immunol. 120, 1781-1782 (abstract).
- Lachmann, P.J. (1979) in The Antigens (Sela, M. Ed.) Vol. 5,

- pp 283-353, Academic Press, N.Y.
- Lachmann, P.J. (1983) *Nature* 305, 473-474.
- Lachmann, P.J. (1984) *Phil. Trans. Soc. R. Soc. Lond. B.* 306, 419-430.
- Lachmann, P.J. and Hugh-Jones, N.C. (1984) *Springer. Semin. Immunopathol.* 7, 12-32.
- Lachmann, P.J. and Hobart, M.J. (1985) *Trends Genetics* 1, 145-150.
- Lambris, J.D. and Muller-Eberhard, H.J. (1986) *Molec. Immunol.* 23, 1237-1242.
- Laskey, R.A. and Mills, A.D. (1977) *FEBS Letts.* 82, 314-316.
- Law, S.K.A., and Levine, R.P. (1977) *Proc. Natl. Acad. Sci. USA* 74, 2701-2705.
- Law, S.K.A., Lichtenberg, N.A., Holcombe, F.H. and Levine, R.P. (1980a) *J. Immunol.* 125, 634-639.
- Law, S.K.A., Lichtenberg, N.A. and Levine, R.P. (1980b) *Proc. Natl. Acad. Sci. USA* 77, 7149-7198.
- Law, S.K.A. (1983) *Ann. N.Y. Acad. Sci.* 421, 246-258.
- Law, S.K.A., Dodds, A.W. and Porter, R.R. (1984a) *EMBO J.* 3, 1819-1823.
- Law, S.K.A., Minich, T.M. and Levine, R.P. (1984b) *Biochemistry* 23, 3267-3272.
- Law, S.K.A. and Gagnon, J. (1985) *Biosci. Rep.* 5, 913-921.
- Levi-Strauss, M., Tosi, m., Steinmetz, M., Klein, J. and Meo, T. (1985) *Proc. Natl. Acad. Sci. USA* 82, 1746-1750.
- Lintin, S. and Reid, K.B.M. (1986) *FEBS Letts.* 204, 77-81.
- Longster, G. and Giles, C.M. (1976) *Vox Sang.* 30, 175-180.
- Lublin, D.M., Lemons, R.S., Beau, M.M.L., Holer, V.M., Tykocinski, M.L., Medof, M.E. and Atkinson, J.P. (1987) *J. Exp. Med.* 165, 1731-1736.
- Lundwall, A., Hellman, U., Eggertsen, G. and Sjogquist, J. (1982) *Molec. Immunol.* 19, 1655-1665.
- Lundwall, A., Wetsel, R.A., Domdey, H., Tack, B.F. and Fey, G.H. (1984) *J. Biol. Chem.* 259, 13851-13856.
- Lundwall, A.B., Wetsel, R.A., Kristensen, T., Whitehead, A.S., Woods, D.E., Ogden, R.C., Colten, H.R. and Tack, B.F. (1985) *J. Biol. Chem.* 260, 2108-2112.

- Maniatis, T., Frisch, E.F. and Sambrook, J. (1982) *Molecular Cloning, a Laboratory Manual.*, Cold Spring Harbour Laboratory Press, N.Y.
- Matthew, W.J., Goldberger, G., Marino, J.T., Einstein, L.P., Gash, D.J. and Colten, H.R. (1982) *Biochem. J.* 204, 839-846.
- Mauff, G., Alper, C.A., Awdeh, Z., Batchelor, J.R., Bertrams, T., Bruun-Petersen, G., Dawkins, R.L., Demant, P., Edwards, J., Grosse-Wilde, H., Hauptmann, G., Klouda, P., Lamn, L., Mollenhauer, E., Nerl, C., Olaisen, B., O'Neill, G., Rittner, C., Ross, M.H., Skanes, V., Tieberg, N. and Wells, L. (1983a) *Immunobiol.* 164, 184-191.
- Mauff, G., Steuer, M. and Bender, K. (1983b) *Hum. Genet.* 64, 186-188.
- Maxam, A.M. and Gilbert, W. (1980) *Methods in Enzymol.* 65, 499-560.
- Medof, M.E., Lida, K., Mold, C. and Nussenzweig, V. (1982) *J. Exp. Med.* 156, 1739-1754.
- Medof, M.E., Kinoshita, T., Silber, R. and Nussenzweig, V. (1985) *Proc. Natl. Acad. Sci. USA* 82, 2980-2984.
- Medof, M.E., Walter, E.I., Roberts, W.L., Haas, R. and Rosenberry, T.L. (1986) *Biochemistry* 25, 6740-6747.
- Medof, M.E., Lublin, D.M., Holers, V.M., Ayers, D.J., Getty, R.R., Leykam, J.F., Atkinson, J.P. and Tykocinski, M.L. (1987) *Proc. Natl. Acad. Sci. USA* 84, 2007-2011.
- Messing, J., Gronenborn, B., Muller-Hill, B., and Hofschneider, P.H. (1977) *Proc. Natl. Acad. Sci. USA* 74, 3642-3646.
- Messing, J., Crea, R. and Seeburg, P.H. (1981) *Nucleic Acid Res.* 9, 309-321.
- Messing, J. and Vierira, J. (1982) *Gene* 19, 269-276.
- Messing, J. (1983) *Methods in Enzymol.* 101, 20-78.
- Middleton, J. and Crookston, M. (1972) *Vox Snag.* 23, 256-261.
- Middleton, J., Crookston, M.C., Falk, J.A., Robson, E.B., Cook, P.J.L., Batchelor, J.R., Bodmer, J., Ferrara, G.B., Festenstein, H., Harris, R., Kissmeyer-Nielsen, F., Lawler, S.D., Saches, J.A. and Wolf, E. (1974). *Tissue Antigens* 4, 366-373.
- Mitomo, K., Fujita, T. and Lida, K. (1987) *J. Exp. Med.* 165, 1424-1429.
- Moon, K.E., Gorski, J.P. and Hugli, T.E. (1981) *J. Biol. Chem.* 256, 8685-8692.

- Morley, B.J. and Campbell, R.D. (1984) EMBO J. 3, 153-157.
- Muller, U., Stephan, D., Philippsen, P. and Steinmetz, M. (1987) EMBO J. 6, 369-373.
- Muller-Eberhard, H.J. (1981) In 'Advance in Immunopathology' pp. 141, ed. Weigle, W.O.
- Muller-Eberhard, H.J. (1986) Ann. Rev. Immunol. 4, 503-528.
- Nagasawa, S. and Stroud, R.M. (1977) Proc. Natl. Acad. Sci. USA 74, 2998-3001.
- Nagasawa, S., Ichihara, C. and Stroud, R.M. (1980) J. Immunol. 7, 578-582.
- Naora, H. and Deacon, N.J. (1982) Proc. Natl. Acad. Sci. USA 79, 6196-6200.
- Nedwin, G., Naylor, S.L., Sakaguchi, A.Y., Smith, D., Jarret-Nedwin, J., Pennica, D., Goeddel, D. and Gray, P.W. (1985) Nucl. Acid Res. 13, 6361-6373.
- Ng, S.C., Rao, A.G., Howard, O.M.Z. and Sodetz, J.M. (1987) Biochemistry 26, 5229-5233.
- Nisonoff, A., Hopper, J.E. and Spring, S.B. (1975) In The antibody molecule, pp 370-388. Academic Press, New York.
- Nonaka, M., Nakayama, K., Yu, D.Y., Shimizu, A. and Takahashi, M. (1985) Immunol. Rev. 87, 81-99.
- Nordhagen, R., Olaisen, B., Teiberg, P. and Gedde-Dahl Jr, T. (1980) J. Immunogenet. 7, 301-306.
- Novotny, J., Handschumacher, M. and Bruccoleri, R.E. (1987) Immunol. Today 8, 26-31.
- O'Neill, G.J., Yang, S.Y., Tegoli, J., Berger, R. and Dupont, B. (1978a) Nature 273, 668-670.
- O'Neill, G.J., Yang, S.Y. and Dupont, B. (1978b) Proc. Natl. Acad. Sci. USA 75, 5165-5169.
- Palsdottir, A., Fossdal, R., Arnason, A., Edwards, J.H. and Jensson, O. (1987) Immunogenetics 25, 299-304.
- Pangburn, M.K. and Muller-Eberhard, H.J. (1980) J. Exp. Med. 152, 1102-1114.
- Parkes, C., Gagnon, J. and Kerr, M. (1983) Biochem. J. 213, 201-219.
- Partanen, J. and Koskimies. (1986) Human Hered. 36, 269- 275.
- Partanenn, J., Koskimies, Ieonen, J. and Knip, M. (1986) Tissue Antigens 27, 291-297.

- Patthy, L. (1987) FEBS Letts. 214, 1-7.
- Pericak-Vance, M.A., Elston, R.C., Spira, T.J. and Band, J. (1982) Am. J. Hum. Genet. 34: 90 (abstr.)
- Petes, T. and Fink, G.R. (1982) Nature 300, 216-217.
- Philips, D.C., Sternberg, M.J.E. and Sutton, B.J. (1983) In Evolution from molecules to men. ed. D.S. Bendall, pp. 145-173 Cambridge University Press, Cambridge.
- Podack, E.R., Tschopp, J. and Muller-Eberhard, H.J. (1982) J. Exp. Med. 156, 268-282.
- Polley, M.J. and Muller-Eberhard, H.J. (1967) J. Exp. Med. 126, 1013-1025.
- Porter, R.R. (1983) Mol. Biol. Med. 1, 161-168.
- Porter, R.R. (1984) CRC Crit. Rev. Biochem. 16, 1-19.
- Porter, R.R. (1985) Immunol. Rev. 87, 7-17.
- Proudfoot, N.J. and Brownlee, G.G. (1976) Nature, 263, 211-214.
- Prydzial, E.L. G. and Isenman, D.E. (1987) J. Biol. Chem. 262, 1519-1525.
- Radloff, R., Bauer, W. and Vinograd, J. (1967) Proc. Natl. Acad. Sci. USA 57, 1514-1521.
- Rao, A.G., Howard, O.M.Z., Ng, S.C., Whitehead, A.S., Colten, H.R., and Sodetz, J.M. (1987) Biochemistry 26, 3556-3564.
- Reid, K.B.M. and Porter, R.R. (1976) Biochem. J. 155, 19-23
- Reid, K.B.M. and Porter, R.R. (1981) Ann. Rev. Biochem. 50, 433-464.
- Reid, K.B.M. (1983) Biochem. Soc. Trans. 11, 1-12.
- Reid, K.B.M. (1985) Biochem. J. 231, 729-735.
- Reid, K.B.M. (1986) Essays in Biochem. 22, 27-68.
- Reid, K.B.M., Bentley, D.R., Campbell, R.D., Chung, L.P., Sim, R.B., Kristensen, T. and Tack, B.F. (1986) Immunol. Today 7, 230-234.
- Reid, K.B.M., Campbell, R.D., Day, A.J., Dodds, A.W., Law, S.K.A., Lintin, S.J., Ripoche, J., Sim, R.B., Wu, L.C. and Yu, C.Y. (1987) Prog. Immunol. (in press).
- Reveille, J.D., Arnett, F.C., Wilson, R.W., Bias, W.B. and McLean, R.H. (1985) Immunogenetics 21, 299-311.
- Rey-Campes, J., Rubinstein, P. and Rodriguez de Cordoba, S. (1987) J. Exp. Med. 166, 246-252.

- Rigby, P.W.J., Dieckmann, M., Rhodes, C. and Berg, P. (1977) *J. Mol. Biol.* 113, 237-251.
- Ripoche, J., Day, A.J., Willis, A.C., Belt, K.T., Campbell, R.D. and Sim, R.B. (1986) *Bioscience Rep.* 6, 65-72.
- Rittner, C., Hauptmann, G., Gross-Wilde, H., Grosshans., Tongio, M.M. and Mayer, S. (1975) In *Histocompatibility Testing*, Munksgaard, Copenhagen, pp 945-953.
- Rittner, C., Giles, C.M., Roos, M.H., Demant, P. and Mollenhauer, E. (1984) *Immunogenetics*, 19, 321-333.
- Rodrigues, N.R., Dunham, I., Yu, C.Y., Carroll, M.C., Porter, R.R. and Campbell, R.D. (1987) *EMBO J.* 6, 1653-1661.
- Rodriguez de Cordoba, S., Lublin, D.M., Rubinstein, P. and Atkinson, J.P. (1985) *J. Exp. Med.* 161, 1189-1195.
- Rogde, S., Olaisen, B., Gedde-Dahl, T. and Teisberg, P. (1985) *Cytogenet. Cell Genet.* 40, 734-735 (Abstr.).
- Roos, M.H., Mollenhauer, E., Demant, P. and Rittner, C. (1982) *Nature* 298, 854-855.
- Roos, M.H., Giles, C.M., Demant, P., Mollenhauer, E. and Rittner, C. (1984) *J. Immunol.* 133, 2634-2640.
- Roos, G.D. and Medof, M.E. (1985) *Adv. Immunol.* 37, 217-267.
- Sakano, H., Rodgers, J.H., Huppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R. and Tonegawa, S. (1979) *Nature* 277, 627-633.
- Sanger, F., Nicklsen, S., and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463-5468.
- Sanger, F. and Coulson, A.R. (1978) *FEBS Letts.* 87, 107-110.
- Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. and Roe, B. (1980) *J. Mol. Biol.* 143, 161-178.
- Schechter, I. (1971) *Ann. N.Y. Acad. Sci.* 190, 394-419.
- Schifferli, J.A., Hauptmann, G. and Paccaud, J.-P. (1987) *FEBS Letts.* 213, 415-418.
- Schneider, P.M., Carroll, M.C., Alper, C.A., Rittner, C., Whitehead, A.S., Yunis, E.J. and Colten, H.R. (1986) *J. Clin. Invest.* 78, 650-657.
- Schendel, D.J., O'Neill, G. and Wank, R. (1984) *Immunogenetics* 20, 23-31.
- Schreiber, R.D. and Muller-Eberhard, H.J. (1974) *J. Exp. Med.* 140, 1324-1335.

- Sepich, D.S., Noonan, D.J. and Ogata, R. (1985) *Proc. Natl. Acad. Sci. USA* 82, 5895-5899.
- Seya, T., Nagasawa, S. and Atkinson, J.P. (1986a) *I. Immunol.* 136, 4152-4156.
- Seya, T., Turner, J.R. and Atkinson, J.P. (1986b) *J. Exp. Med.* 163, 837-855.
- Sharp, P.A., Sugden, B. and Sambrook, J. (1973) *Biochemistry* 12, 3055-3063.
- Sharp, P.A. (1981) *Cell* 23, 643-646.
- Sim, R.B. and Sim, E. (1981) *Biochem. J.* 193, 129-141.
- Sim, E., Wood, A.B., Hsiung, L.M. and Sim, R.B. (1981) *FEBS Letts.* 132, 55-60.
- Sim, E. and Cross, S. (1986) *Biochem. J.* 239, 763-767.
- Sim, R.B., Malhotra, V., Ripoche, J., Day, A.J., Micklem, K.J. and Sim, E. (1986) *Biochem. Soc. Symp.* 51, 83-96.
- Sim, R.B., Malhotra, V., Day, A.J. and Erdei, A. (1987) *Immunol. Letts.* 14, 183-190.
- Singer, M.F. (1982) *Cell*, 28, 433-434.
- Skanes, V.M., Larsen, B. and Giles, C.M. (1985a) *Immunogenetics* 22, 609-616.
- Skanes, V., Larson, B., Sampon-Murphy, L. and Farid, N.R. (1985b) *Clin. Invest. Med.* 8, 126-132.
- Smith, C.A., Vogel, C.W., and Muller-Eberhard, H.J. (1984a) *J. Exp. Med.* 159, 324-329.
- Smith, C.A., Pangburn, M.K., Vogel, C.W. and Muller-Eberhard, H.J. (1984b) *J. Biol. Chem.* 259, 4582-4588.
- Sottrup-Jensen, L., Petersen, T.E. and Magnusson, S. (1980) *FEBS Lett.* 121, 275-279.
- Sottrup-Jensen, L., Folkersen, J., Kristensen, T. and Tack, B.F. (1984) *Proc. Natl. Acad. Sci. USA* 81, 7353.
- Southern, E.M. (1975) *J. Mol. Biol.* 98, 503-517.
- Southern, E.M. (1979) *Methods in Enzymol.* 68, 152-176.
- Sproat, B.S. and Gait, M.J. (1984) In Gait M.J. (eds), *Oligonucleotide synthesis*, IRL Press, Oxford, pp 83-115.
- Staden, R. (1982a) *Nucleic Acid Res.* 10, 4731-4751.
- Staden, R. (1982b) *Nucleic Acid Res.* 10, 2951-2961.

- Staden, R. (1984) *Nucl. Acid Res.* 12, 521-538.
- Stanley, K.K., Kocher, H.-P., Luzio, J.P., Jackson, P. and Tschopp, J. (1985) *EMBO J.* 4, 375-382.
- Stein, J.P., Catterall, J.F., Kristo, P., Means, A.R. and O'Malley, B.W. (1980) *Cell* 21, 681-687.
- Steinmetz, M. (1984) *Trends Biochem.* 9, 224-226.
- Tack, B.F., Janatova, J., Thomas, M.L., Harrison, R.A. and Hammer, C.H. (1981) *Methods Enzymol.* 80, 64.
- Tack, B.F. (1985) in Muller-Eberhard M.J. and Meischer, P.J. (eds), *Complement*, Springer-Verlag, Berlin, pp 49-72.
- Takata, Y., Tamura, N. and Fujita, T. (1984) *J. Immunol.* 132, 855-858.
- Takata, Y., Kinoshita, T., Kozono, H., Takeda, J., Tanaka, E., Hong, K. and Inoue, K. (1987) *J. Exp. Med.* 165, 1494-1507.
- Teisberg, P., Akesson, I., Olaisen, B., Gedde-Dahl, T. Jr. and Thorsby, E. (1976) *Nature* 264, 253-254.
- Thielens, N.M. and Colomb, M.G. (1986) *Eur. J. Immunol.* 16, 617-622.
- Tilley, C.A., Romans, D.G. and Crookston, M.C. (1978) *Nature* 276, 713-715.
- Tokunaga, K., Omoto, K., Akaza, T., Akiyama, N., Amemiya, H., Naito, S., Sasazuki, T., Satoh, H. and Juji, T. (1985) *Immunogenetics* 22, 359-365.
- Tosi, M., Duponchel, C., Bourgarel, P., Colomb, M. and Meo, T. (1986) *Gene* 42, 265-272.
- Tschopp, J., Muller-Eberhard, H.J. and Podack, E.R. (1982) *Nature* 298, 534-538.
- Tschopp, J., Masson, D. and Stanley, K.K. (1986) *Nature* 322, 831-834.
- Tse, A.G.D., Barclay, A.N., Watts, A. and Williams, A.F. (1985) *Science* 230, 1003-1008.
- Uring-Lambert, B., Goetz, J., Tongio, M.M., Mayer, S. and Hauptmann, G. (1984) *Tissue Antigens* 24, 70-72.
- Van Regenmortel, M.H.V. (1987) *Trends Biochem. Sci.* 12, 237-240.
- Villiers, M.-B., Thielens, N.M. and Colomb, M.G. (1985a) *Biochem. J.* (1985a).
- Villiers, M.-B., Thielens, N.M., Villiers, C.L. and Colomb,

- M.G. (1985b) *Eur. J. Immunol.* 15, 941-945.
- Vogt, W., Schmidt, G., Buttlar, B.V. and Dieminger, L. (1978) *Immunology* 34, 29-40.
- Wahl, G.M., Stern, M. and Stark, G.R. (1979) *Proc. Natl. Acad. Sci. USA* 76, 3683-3687.
- Weiher, H., Konig, M. and Gruss, P. (1983) *Science* 219, 626-631.
- Weis, J.J., Fearon, D.T., Klickstein, L.B., Wong, W.W., Richard, S.A., Kops, A.D.B., Smith, J.A. and Weis, J.H. (1986) *Proc. Natl. Acad. Sci. USA* 83, 5639-5643.
- Weis, J.H., Morton, C.C., Bruns, G.A.P., Weis, J.J., Klickstein, L.B., Wong, W.W. and Fearon, D.T. (1987) *J. Immunol.* 138, 312-315.
- Weiss, E., Golden, L., Zakut, R., Mellor, A., Fahrner, K., Krist, S. and Flavell, R.A. (1983a) *EMBO J.* 2, 453-462.
- Weiss, E.H., Mellor, A., Golden, L., Fahrer, K., Simpson, E., Hurst, J. and Flavell, R.A. (1983b) *Nature* 301, 671-673.
- Wetlaufer, D.B. (1981) *Adv. Prot. Chem.* 34, 61-92.
- Wetsel, R.A., Lundwall, A., Davidson, F., Tack, B.F. and Fey, G.H. (1984) *J. Biol. Chem.*, 259, 13857-13862.
- White, P.C., Grossberger, D., Onufer, B.J., Chaplin, D.D., New, M.I., Dupont, B. and Strominger, J.L. (1985) *Proc. Natl. Acad. Sci. USA* 82, 1089-1093.
- Williams, A.F. (1984) *Immunol. Today* 5, 219-221.
- Williams, S.A., Slatko, B.E., Moran, L.S. and DeSimone, S.M. (1986) *Biotechniques* 4, 138-147.
- Winter, G., Fields, S., Gait, M.J. and Brownlee, G.G. (1981) *Nucleic Acid Res.* 9, 237-245.
- Woods, D.E., Markham, A.F., Ricker, A.T., Goldberger, G. and Colten, H.R. (1982) *Proc. Natl. Acad. Sci. USA* 79, 5661-5665.
- Wu, L.C., Morley, B.J. and Campbell, R.D. (1987) *Cell* 48, 331-342.
- Wu, R. and Taylor, E. (1971) *J. Mol. Biol.* 57, 491-551.
- Young, J. D-E., Cohn, Z.A. and Podack, E.R. (1986) *Science* 233, 184-190.
- Yu, C.Y., Belt, K.T., Giles, C.M., Campbell, R.D. and Porter, R.R. (1986) *EMBO J.* 5, 2873-2881.
- Yu, C.Y. and Campbell, R.D. (1987) *Immunogenetics* 25, 383-390.

Ziccardi, R.J. (1986) J. Immunol. 136, 3378-3385.

Ziccardi, R.J., Dahlback, B. and Muller-Eberhard, H.J.
(1984) J. Biol. Chem. 259, 13674-13679.

Appendix I

Genomic DNA and derived amino acid sequences of the C4A 3 gene from Cos 3A3

The DNA sequence of the C4A gene was determined as described in Chapter 3. The exon-intron structure was deduced by comparing the genomic DNA sequence of C4 with published C4 cDNA sequences (Belt et al., 1984). Numbering of amino acids is after Belt et al., (1984). The extra amino acids detected by genomic DNA sequences are denoted by +. Numbering of the nucleotides are defined by (A) the position of the major transcriptional start site (marked by asterisk) as nucleotide A1, which applies to the sequence 5' to 6-7 kb intron; and (B) the first nucleotide of the recognition site of the restriction enzyme Stu I located 5' to exon 11, as nucleotide B1, which applies to the DNA sequence 3' to 6-7 kb intron. The C4 gene contains 42 exons and they altogether encodes 1745 amino acid residues. The thiolester site (overlined) is encoded by exon 25. The residues determining the C4A/C4B isotypicity and the Rg/Ch antigenic determinants are marked by dots over the amino acid sequences. The transcriptional enhancer core consensus is underlined. The poly-A signal is marked by dots over the nucleotide sequence. The poly-A site is marked by an asterisk.

A

TCCTTTATTTACTGTGACGTGTGTTTGGGTGCTTGTATGAGGTGAGGTCTG -541

AGCTTTGCTCATTGAGAGGGAAGGGGTGAGGGTCACTCTGACGTTGAGGCCATTG -481

TCCCTCCTGGAATGCTGAGGCTGACCTAATCTCCTAACCACGGAATTTCTGTTAGGGC -421

CTAAAAAGCAAAAGCCTAGTATAGTTCAATTTGTGTTGGAATGAAAGTAAGAGACAGT -361

GTCTTAGAAGCCTGTCTTTTGTGAGGGCTTTAATATCCTGTACTCTGGGCCAT -301

GTTGGGCCCTTGTACGCCCAAGTATACATGAGCTTGTGTGACCTATACCTGATACAG -241

TATACCTGCTAGGGGAGGCTGCTCAGGCACTGGAATGAGAGGATTAACGGGAAAGACA -181

GGGTTATTTCTGGGCCAAGATTGAGATTTCCCATGGACACCAAGGTGTCCGGGGTGGC -121

enhancer core

CCACAACCTGTGGCTGAGGCCAGTTGCACTTCTTGGCTGTACGTTGTTCCAGCTTA -61

CAP site *

GCTGGGCTGGGGAGGAGCAAGGTCCAGAGTCAACTCTGCCCCGAGGCTAGCTTGCCCA 1

(EXON 1) -19

M R L

GAGGTAAGCAGACAGACAGGATCTAACCTCTTGGATCCTCCAGCCATGAGGCTGC 61

1 3

L W G L I W A S S F F T L S L Q K P R

TCTGGGGCTGATCTGGGCATCCAGCTTCTTCACTTATCTCTGAGAAAGCCAGGCTCT 121

GGAGGGGGGATGCTGGGTGCTTGGATTGGGGCAGGGCTGGCATCGGGACCCGATTGAGGA 181

GTGAGGGAGAGCAGGGGTGAGGCTGAGAGCGAAGTCTGACTGCTGATCCTGTCTGTTT 241

4 (EXON 2)

L L L F S P S V V H L G V P L S V

TCCCCAGGTTGCTTGTCTCTCTCTCTGTTGATCTGGGGTCCCCCTATCGGTGG 301

G V Q L Q D V P R G Q V V K G S V F L R

GGGTGCACTCCAGGATGTGCCCCGAGGACAGGTAGTGAAAGGATCAGTGTCTCTGAGAA 361

N P S R N N V P C S P K V D F T L S S E

ACCCATCTCGTAATATGTCCCTGCTCCCCAAGGTGGACTTACCCCTTAGCTCAGAAA 421

69

R D F A L L S L Q

GAGACTTGCACCTCTGATCTCCAGGTAAACAGACCCCATGCCCCCTGCTGCTTGTG 481

GGGGCTCTGCCCCGTTCCCATCTGATCTTGTAGTCTGTCATCTGCCCCACTGGCCT 541

CCTCCCCCTCTGCTTCCCAACCTGGCATCTCTCTTCCAGGTTTCTCCCTTGGTCTCTGT 601

CCTTTTGTGCTGCTGCTCTTGTGCTGTGACCCGCTCCCTCTCCCTCTCCCTCTCTGA 661

70 (EXON 3)

V P L K D A K S C G L H Q L L R G P E

CAGGTGCCCTTGAAAGATGCGAAGAGCTGTGGCTCCATCAACTCTCAGAGGCCCTGAG 721

V Q L V A H S P W L K D S L S R T T H I

GTCCAGCTGGTGGCCCATTCGCCATGGCTAAGGACTCTCTGTCCAGAACGACAAACATC 781

Q G I N L L F S S R R G H L F L Q T D Q

CAGGGTATCAACCTGCTCTTCTCTCTCTGCCCCGGGGCACCTCTTTTTCAGACGGACCAG 841

136

P I Y N P G Q R

CCCATTACAACTTGGCCAGCGGGGTGAGTCTAGCCCCAGGGCTCAACCTTTAAGCC 901

CCTCCGAGCCCTCTCAGGATGAGTTTGTGCCCCCTAAGTGAGATAACCTGAAAGAAAGT 961

GCCACACAGAAAGGGGTGCTTAGGAACATTTGTCCTGCTCCCTCTGTGGAGTTTGACC 1021

137

V R

CACCCTCCCTTGCACATGGACCCCTGCTCAGCTCTCTCTCTCCCTCCACTCCAGTTGAGT 1081

(EXON 4)

Y R V F A L D Q K M R P S T D T I T V M

ACCGGGTCTTTGCTCTGGATCAGAGATGCGCCGAGCACTGACACCATCACAGTCATGG 1141

160

V E

TGGAGGTGAGTCCCCGACCTCTGGCCTTCTGTCTCTGGCCACTGATGTGACCTCTGCCT 1201

161 (EXON 5)
N S H G L R V R K K E V
GTGAGCTACTTCTCCCTTGCAGAACTCTCACGGCCTCCGCGTGCAGAAAGAGAGGTGT 1261

190
Y M P S S I F Q D D F V I P D I S E
ACATGCCCTCCTCATCTTCCAGGATGACTTTGTATCCAGACATCTCAGAGTGAAGC 1321

TCCCAATGTGGGGCTGCCCCAAGCTACACACCCCAATTCTGTATAGGCTCTCCACCT 1381

CCCACACAGAGGACGTCCTCCAGATGCCCTGACCTCAGCCTCCTGAGGCTCTGTTAA 1441

CCCCCAGCTCTCTTCCAGGAGAGCAGGCTGCTGCTCTCCGTGCCCCACTGTACAG 1501

191 (EXON 6) 196
P G T W K I
ATGGGCTGAGCCCTTCTTGTCCATTCTCAGGCCAGGACCTGGAAGATCGTCAGAGGA 1561

CGGAGGCAATTGCGCACTGCGGACGATTATGGTGATTTTTCGATACCGTCGGTGGATT 1621

TTTCATAACGAATGCTCTGAAACCTTCGAGAGCTGCTGCCATTCAACCGGTGCTAAA 1681

+ 197
I S A
GCATTGCTTCATCAGGATAATCATATTGTGTCTTTAGTCAGAAACGCAAGATCTCAGC 1741

(EXON 7) 217
R F S D G L E S N S S T Q F E V K K Y
CCGATTCTCAGATGGCTGGAATCCAACAGCAGCAGCCAGTTTGAAGTGAAGAAATATGG 1801

TGAGAGCTGGAACTGGAGGGACAGGCACTGCTTCTGAAGGAATAGGGTGGAGGG 1861

AGAGGTACTGGGAGCAGCTCAGGGCAAGGAGATATGGGTGCCACAGCCCTGAGCAGAGGG 1921

218 (EXON 8)
V L P N
GAGTCTTTGAGCTGGAGTCTGACCTGCCTATCCCTTACCCCTGGGTGAGTCTTCCCAAC 1981

F E V K I T P G K P Y I L T V P G H L D
TTTGAAGTGAAGATCAGCCCTGGAAGGCCCTACATCTGACGGTGCCAGGCCATCTTGAT 2041

250
E M Q L D I Q A R
GAATGCAATTAGACATCCAAGGCCAGGTAATACCTCCCTCCCACTCTGCCCACCAACA 2101

CCGGGTCTGCTCCCTACTCAGTATGAATGGGCTCCTGCTCCCTGCCCCTGGGGCATT 2161

251 (EXON 9)
Y I Y Q K
TTCCCCCAGCCCTTGGCCACCCCTCTTCTCTGCCACGACAGGTACATCTATGGGAAG 2221

P V Q Q V A Y V R F G L L D E D G K K T
CCAGTGCAAGGGGTGCAATATGTGCGCTTTGGGCTCCTAGATGAGGATGGTAAGAGACT 2281

285
F F R G L E S Q T K
TTCTTTTGGGGCTGGAGAGTCAGACCAAGGTAAGAGGAGAGATAGGGGCTGGGGAGGGG 2341

AAGGGGCAAGGGAGGTGAGGTGGGAGACTCAGTCTCACCTATGCTCTTTCTTTCTAT 2401

286 (EXON 10)
L V M G Q S H I S L S K A E F Q D A
GCCCCAGCTGGTGAATGACAGAGCCACATTTCCCTCTCAAGGGCAGGTTCAGGAGCGC 2461

L E K L N M G I T D L Q G L R L Y V A A
CCTGAGAGAGCTGAATATGGGCATTACTGACCTCAGGGGCTGCGCCTCTACGTTGCTGC 2521

329
A I I E Y P G
AGCCATCATTGATATCAAGGTGGGTGACTTTCCCTTATTGTACCCCAAGCCCTTGCCT 2581

~7kb INTRON

~7kb INTRON
AAGCCTCTTGTGGCTTTCCACAACCTATTGTCCCATATTTTATGGCCAATTTATACAGG 60

CACCCACAGGTCCTTTTCCACACAGACAGGAATACGGCAGCCTGTGCCCTGGGAGCT 120

CACTGTCTTGTGGGAGGGAACTCAAGCCACTCCCACTTGTCTCTGTCCCTCTCT 180

330 (EXON 11)
G G E M E E
TCTTGGGCTCTGTCCCCACCTCTCTCTGTCTTGTCTTGTGAGGTGGGGAGATGGAGGA 240

A E L T S W Y F V S S P F S L D L S K T
GGCAGAGCTCACATCCTGTATTTGTGTCTCTCTCTTCTCTTGGATCTTACAGAGAC 300

369
K R H L V P G A P F L L Q
CAAGCAGACCTTGTGCTGGGGCCCCCTTCTGCTGCAAGTTTCTTCAAGAGGGAGAGG 360

ATGAGTAGGGAGGATGTGTAGTTAGGAGGGCTCAGGGTCTGACCACTCTCTTTTGCCTG 420

(B)

370 (EXON 12)
A L V R E M S G S P A S G I
CCCTCCTTTACCTGCCTAGGCCTTGCTCCGTGAGATGTCAAGCTCCCCAGCTTCTGGCAT 480

P V K V S A T V S S P G S V P E V Q D I
TCCTGTCAAAATTTCTGCCACGCTGCTTCTCCTGGGTCTGTTCTGAAGTCCAAGACAT 540

Q Q N T D G S G Q V S I P I I I P Q T I
TCAGCAAAACACAGACGGGAGCGGCCAAGTCAGCATTCCAATAATTATCCCTCAGACCAT 600

428
S E L Q L S
CTCAGAGCTGCAAGCTCTCAGTAAGACTCCTCAGACCCCTGGAAGATGCTGAGGGAAGAGAG 660

AGGAGGCTGAGCTGAGGTCCCAAGATCCATGCTGACTTGGGAGGAAGGTGAGGACT 720

429 (EXON 13)
V S A G S
TGGCTCTGAGCTACTACCTATTGACACCTGACCCCTCTCCAGGTATCTGAGGCTCCC 780

P H P A I A R L T V A A P P S G S P S F
CACATCCAGCAGATAGCCAGGCTCACTGTGAGAGCCCACTTCAGGAGGCCCCGAGTTTC 840

L S I E R P D S R P P R V G D T L N L N
TGTCTATTGAGCGGCCGATTCTCGACCTCCTGTTGGGAGACTCTGAGCTGAGCT 900

489
L R A V G S G A T F S H Y Y Y M
TGCAGCCCTGAGCAATGAGGACCTTTTCTCATTACTACTACATGCTGATGAGCT 960

GAGGAGTCACGAGGCTGAGGTGAGGAGAGAGCCCTCTGAGTGGGCTGAGGAGGTTT 1020

AGGCTGAGGCTGCTCCATGAGAGAGGCAACCACTCTTGTCCCTCCCATTTCTTGGCCAGA 1080

490 (EXON 14)
I L S R G Q I V F M N R E P K R T L T S
TCCTATCCGAGGAGCAGATGCTGTTTCAATGAGAGCCCAAGAGGACCTGACCTCGG 1140

V S V F V D H H L A P S F Y F V A F Y Y
TCTCGGTGTTTGTGAGCATCACCTGGCACCCTCCTTCTACTTTGTGGCTTCTACTACC 1200

H G D H P V A N S L R V D V Q A G A C E
ATGAGAGCCACCCAGTGGCCAACTCCCTGCGAGTGGATGTCCAGGCTGGGAGCTGCGAGG 1260

551
G K
GCAAGGTGACCAGGCTCAAGAGAGATGACACTTGTGCGAGAGGAGGTTGAGGACAGGCTG 1320

ATTGCCAAGAGGATGAGATTTAGCTTGGGAGCAGTGAAGATACGGGACTGAAGGAGAGC 1380

552 (EXON 15)
L E L S V D G
TCTCCCACTCTGACCAGCCCACTGACCCCTGCAAGCTGAGAGCTCAGCTGAGAGCT 1440

A K Q Y R N G E S V K L H L E T D S L A
GCCAAGCAGTACCGAGAGGAGGAGTCCGTGAAGCTCCACTTGAAGACCACTCCCTAGCC 1500

L V A L G A L D T A L Y A A G S K S H K
CTGGTGGCTGAGAGGCTTGGAGCAGCTCTGTATGCTGAGAGCAGCAAGTCCACAGG 1560

604
P L N M S K
CCCTCAACATGAGCAAGGTTTGTCCAGACCCCTCTCCACAAGCTCTCTACCCCTCCATGG 1620

CTCATCCCCCTGCTTCCCTGAGCCTTGGGAGCAGCCCTGGATCCCACTGAGGCTCCCA 1680

CAGTCTCTTCCCACTTGGCCCTGTGGTCTCCATCTCCTGAGCTCTATCTTCTCTATC 1740

CCCCATGTGCTGCTCTCACCTGTGCGAGTGTCTCAGTCTGCTCCCTCAGCCACACTT 1800

605 (EXON 16)
V F E A M N S Y D L
GGCTCTAGCATTTCTGCTTTCTTGAAGCTTGAAGCTATGAGAGCTATGAGCTC 1860

G C G P G G G D S A L Q V F Q A A G L A
GGCTGTGGTCTGAGGAGTGGGAGAGTGGCTTCAAGTGTTCAGGAGCAGGAGCTGGCC 1920

646
F S D Q D Q W T L S R K
TTTTCTGATGAGAGCAAGTGGACCTTATCCAGAGAGAGTGAAGACAGAGAGAGAGAG 1980

GTGGGTGGCGGAGATAGGGAGAGAGAGAGAGGCTGAGGAGACCACTGAGAGAGTC 2040

CAGGAGAGAGGCTGAGCAGGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG 2100

647 (EXON 17)
R L S C P K E K T T R K
GGACTGAGCCTTTCTCTCTACAGGACTAGCTGTCCAGGAGAGAGAGAGAGAGAGAG 2160

672
K R N V N F Q K A I N E K
AAGAGAGAGCTGAAGTTCCAAAGGAGATTAATGAGAGATGTGAGTTGAGGAGCTAGG 2220

[illegible]

933 (EXON 24)
D H R G R T L E I P G N S D P N M I P D
ACCACCGAGGCCGACCTTGGAATACCTGGCACTCTGATCCCAATATGATCCCTGATG 4200

962
G D F N S Y V R V T
GGGACTTTAAGCTACGTCAAGGTTACAGGTGGGAGTGGCCTTTAGTCCCTTCCCACTG 4260
GCCACCTTCGGATTCTGTGGGACCTGTGGATCCCTGCTTGGTCCCACTCCCCGTGAGCC 4320

963 (EXON 25)
A S D P L
TCTGACACAGAGTCTCAGACCTCCACCTCTCCCTCCCATGTAGCTCAGATCCATTGG 4380

D T L G S E S A L S P G G V A S L L R L
ACACTTTAGGCTCTGAGGGGGCCTTGTACCAAGAGGCGTGGCCTCCCTCTTGAAGGCTTC 4440

Thioester
P R G C G E Q T M I Y L A P T L A A S R
CTCGAGGCTGTGGGAGCAACCATGATCTACTTGGCTCCGACACTGGCTGCTTCCCGCT 4500

Y L D K T E Q W S T L P P E T K D H A V
ACCTGGACAAGACAGAGCAAGTGGAGCACTGCTCTCCGAGACCAAGGACCACGCGGTGG 4560

1032
D L I Q K
ATCTGATCCAGAAAGTTCTGGGTGCAAGGCAAGCAAGGAGGGGGCCAGGAAAGGACAG 4620
TTCTGGAAAGATGGACAGCCCAGGAGGCTTACAGAGGGAAAGAAAGGGGGCCCTGATGAG 4680
GATGGGAGCATGGCCTTGGGCTCAACAGCAGAAAGGAGTGTACCTGAGCGGCCAC 4740

1033 (EXON 26)
Y M R I Q Q F R K A D G S Y A
CTCTCCTCTCAAGGCTACATGCGATCCAGCAGTTTCGGAGGGCGGATGGTTCTATGC 4800

1058
A W L S R D S S T W
GGCTTGGTTGTACGAGGACAGCAGCCTGGTGAAGCTTGGGAGAGTGGTTCCAGGGTTCT 4860
GAGGGGGTCAGGGCIGGGGAGGGGGTGGGACAGAGCTGGTATGATGGGAGGGTGGATAAC 4920
CAGGCACCTGGGGGCGTGGGCATATGAGAAACAAATCCTTATCCCCAACCCCTCCTTTC 4980

1059 (EXON 27)
L T A F V L K V L S L A Q E Q V
TGCCCTCCAGGCTCACAGCCTTTGTGTGAAGGCTCTGAGTTTGGCCAGGAGCAGGTAG 5040

G S S P E K L Q E T S N N L L S Q Q Q A
GAGGCTCGCCTGAGAAACTGAGGAGACATCTAAGTGGCTTCTGTCCAGCAGCAGGCTG 5100

1110
D G S F Q D P C P V L D R S M Q
ACGGCTCGTTCCAGGACCCCTGTCCAGTGTAGACAGGAGCATGCAAGTGCAGGCTGCT 5160
GGGGCTGGCCGAGAGCGCCTGTGAGAGGACTCTCTTTGCCCTTCCCCCTCCTGTTTG 5220

1111 (EXON 28)
G G L V G N D E T V A L T
ACATCTTTTCTCCCTTACTAGGGGGGTTTGGTGGGCAATGATGAGACTGTGGCACTCAC 5280

A F V T I A L H H G L A V F Q D E S A E
AGCCTTTGTGACCATCGCCCTTCATCATGGGCTGGCCGCTTCCAGGATGAGGGTGCAGA 5340

1149
P L K Q R V
GCCATTGAAGCAGAGAGTGGTAAGTTCAAGTGGCGTTTCTGCCCTCTGCTGGCCCCAGCT 5400
CTCTCCCTTTTCTCAGGAACCCAGGGGTCAGGGCCAGAGCCCTCCTCCCGTTTCTT 5460

1150 (EXON 29)
E A S I S K A H S F L G E K A S A G L
CCAAGAAAGCTCCATCTCAAGGCAACTCATTTTGGGGGAGAAAGCAAGTGTGGGCT 5520

L G A H A A A I T A Y A L T L T K A P V
CCTGGGTGCCACGCAAGCTGCCATCACGGCTATGCCCTGACACTGACCAAGGGCGCTGT 5580

1206
D L L G V A H N N L M A M A Q E T G
GGACCTGCTCGGTGTTGCCACACCAACCTCATGGCAATGGCCAGGGAGACTGGAGGTGA 5640
GGGGTGAAGGCGCTCTGGAGTGAAGCTGAAGGCCCAAGGGAGCTTAGGATCCTGAAGTGGC 5700
CAGAGGGAGAGGCTGGATGAAGACTCAAGGGAGGAATGAAGTTATAGCAAGGGTGGGTT 5760
GGGGGAGACTCAAGAGAGCCAGCAGGGGGTGGCTAAGGGCCAGGGGACCAAGGCTCTTCT 5820

1207 (EXON 30)
D N L Y W G S
CCCTGCTTCCTGTTTACTTGTGTCTCCCTTCACTTTTCAGATAACCTGTACTGGGGCTC 5880
V T G S Q S N A V S P T P A P R N P S D
AGTCACTGGTTCTCAGAGCAATGCCGTGTGCCCCACCCGGCTCCTGCAACCCATCCGA 5940
P M P Q A P A L W I E T T A Y A L L H L
CCCCATGCCCCAGGCCCCAGCCCTGTGGATTGAACACACAGCCTACGCCCTGCTGCACCT 6000
L L H E G K A E M A D Q A A A W L T R Q
CCTGCTTCACGAGGGCAAGCAGAGATGGCAGACCAAGCTGCGGCTGCTCAGCCGTC 6060
1284
G S F Q G G F R S T Q
GGGCACTTCCAGGGGGATTCCGCAATACCCAAATAGGGGGCCGTCCCCGGGCTCTGGGG 6120
1285
D
GGGGTGGGTAGTCTCAGACCAAGGGCTTGCTTGAGTCTGGCTCAACCTCCCTAGGACA 6180
(EXON 31)
T V I A L D A L S A Y N I A S H T T E E
CGGTGATTGCCCTGGATGCCCTGTCTGCTACTGGATTGCTCCACACCACTGAGGAGA 6240
R G L N V T L S S T G R N G F K S H A L
GGGGTCTCAATGTACTCTCAGCTCCACAGGCCGGAATGGGTTCAAGTCCCACGCGCTGC 6300
1340
Q L N H R Q I R G L E E E L Q
AGCTGACCAACC8CCAGATTGCGGCTGGAGGAGGAGCTGCAAGTGAACCACTCCCTGG 6360
TGAACCACTCCCTCGCTGGGTAGCCAGGACACCTGGGCTCTGTGGCCAGGCCAGAGCC 6420
GTCCCCACCTCCACCCGTGAATCCCCGACGCACTTCTTCTG66GTCTTC6666AA 6480
GACTGACTTCTGGCTGCGTGACCTGGAGCTCTGAGCTTCAGTTTTCTCACTTGTAGAGT 6540
AACATACACAGAGTTCACTTACAGGGCTGTTAGAGGCTGAAGTGAGATAATTCATGTG 6600
CTGGTATAAATTTGTGGAATGTGAGGTGGGGAGAGGAGGTGGGGCTGTTTTGAGGAA 6660
GGAGATAAGTTATTGGAGCCGCAAAACAGGTTTGCTTGTGCTTCTAACATCGCCTTC 6720
1341 (EXON 32)
F S L G S K I N V K V G G N
CCTTTTCTGTTGCTGAAGTTTTCTTGGGCA8CAAGATCAATGTGAAGGTGGGAGGAA 6780
1360
S K G T L K
CAGCAAGGGAACCTGAGGTTGAGGGCCAGGGAGGGGTGGGGCCAGGCACTGGTGGAGG 6840
AGAGGGGTGTGGAGTGAGAGGGCTGTGGGCA8AGGCAATGGTCCGGGGAAGGAGGAGAC 6900
ACCTCA666TTGGTGTCCGTTGCTTCCGCTG66GTGTTTTTCCCCCTGCTTGTCTGC 6960
TTGCTCTCCCCATCTCTGGGTACCTGTGTGTTTCTTTACCCGCTCA8GTGCTGGTGGCTC 7020
CGAATCCCACTCCTCA8CCAGGCTCTTCCCTGAACCATGGGCCCCACTC8TCCCACTC 7080
CCACAGCAGCTCAGACGAGGCAATGTCCAAAGCCCTTCTTCAATCTGTGTCTCTGTCTG 7140
GCTGGTGGGAGCCCTCCCA8CAGGAGCCCA8CCTACTCTAGAGGCGGTGTTAGTGG 7200
CCCCCTCTCCAA8GCTGTCTTATGTCCCTAGTGACTCCTCTGTCTCCCTGCTGGCT 7260
GTGGCCCTTGGTGTGATCTAGATTCTGTGCTGAGACGGCTTCTCCCTACCTGGAAC 7320
TTCTCTCTACCTCCTGTCTCCCTGTCTGATCCACTGTCCACAGGCA8TGACACTGACC 7380
TTCCAAAGCCCCAGCCAGATCAGCCTTGGGGAAGGTA8TCCCCGCTGCCCCAGGCTC 7440
AGATGGCTGGGCTCTGCCCCACCTCCGGCAGACAGCTCTCCTTGTCTACACAGATCC 7500
CCTTGCTTTCTGTCTCTTCCCTGCTTCTTGGCCACAG8ACAAGCTCTTCTTCTCCTT 7560
CAAGCCTTGGCAGAA8GCTTTCTGAGCTTTTCA8TCCAGCCTCTTCCAGACAGTCT 7620
GGAGTGTGCTCTGGGGGAGGCCCCCTGCTTCTTACCTCTCTGTCTGCTGCTGACGCT 7680
TGTGGCGAATGTGGTGGCACTGCTGTGTGTGGACTGTGCA8TGACGGGAGGGAAGAGGG 7740
CTGAAGGCTCAATCTGTAGCCAGGGAGATGCCCTTAGGATGGACCA8AGAGGTC 7800
1361 (EXON 33)
V L R T
TGTGGCTCACATGTCCACGCTCTCTCCCTGCCCCCTTGTGAGCCAGGTCCTTCGTACC 7860
Y N V L D M K N T T C Q D L Q I E V T V
TACAATGTCTGGACATGAAGAACACGACCTGCCAGGACCTACAGATAGAGGTGACAGTC 7920
1392
K G H V E Y T
AAAGGCCACGTCA8TACAC8AGTGA8TGTGGGGGTGGGA8GCTTGGGGCCAGGCA8G 7980

1393
M
8040
(EXDN 34)
E A N E D Y E D Y E Y D E L P A K D D P
8100
D A P L Q P V T P L Q L F E G R R N R R
8160
R R E A P K V V E E Q E S R V H Y T V C
8220
1451
I W
8280
8340
1452 (EXDN 35)
R N S K V G L S G M A I A D V T L L S G
8400
1481
F H A L R A D L E K
8460
8520
1482 (EXDN 36)
L T S L S D R Y V S H F E T E G P H V L
8580
1506
L Y F O S
8640
8700
8760
8820
8880
8940
9000
9060
9120
9180
9240
9300
9360
9420
9480
9540
9600
9660
9720
9780
9840
9900
9960
10020
10080
1507 (EXDN 37)
V P T S R E C V G F E A V Q E V P
10140
1540
V G L V Q P A S A T L Y D Y Y N P
10200
10260
10320

1541 (EXON 38)
 E R R C S V F Y
 TGGAGGGGCTGTGCCCTAACTCCTCTCTGTTCTCCAGAGCGCAGATGTTCTGTGTTTA 10380

G A P S K S R L L A T L C S A E V C Q C
 CGGGGCACCAAGTAAAGACAGACTCTTGCCACCTTGTTCTGCTGAGTCTGCCAGTG 10440

1570
 A E
 TGCTGAGGGTGAAGCTGAAGGCTGAGGCGAGGCAATGAGAGCGAGATGAGCGAGGCCCC 10500

1571 (EXON 39)
 G K C P R Q R R A
 CCCACACTGTCTGATGAGTTCCCAACTTCAGGGAAGTGCCCTGCGCAGCTGCGGCC 10560

L E R G L Q D E D G Y R H K F A C Y Y P
 TGGAGCGGGTCTGCAAGGACGAGGATGGCTACAGGATGAAGTTTGCTGCTACTACCCCC 10620

1603
 R V E Y
 GTGTGAGGTACGCTCACTCTCCACCGAGGCGCTGCTGACCTCCCTCGGGGACCGGC 10680
 CGTTTTGGTCTCTCTGGGTTGTAGCTGCTCCTTTATCATGGTCATGCACGACCTGT 10740
 TTGCTCTGACACCACTTCTACCTCTCAGCCTCAAGTAACCTACCTTTCCCTCTCT 10800

1604 (EXON 40)
 G F Q V K V L R E D S R A A F
 CCTACCCCCCTCTTAGGCTTCCAGGTTAGGTTCTCCGAGAGACAGCAGAGCTGCTTTC 10860

1631
 R L F E T K I T Q V L H F
 CGCCTCTTTGAGACCAAGATCACCAAGTCTGCACTTCAGTATGAAGCAAGCCGAGAG 10920
 GCGGGCAGGGCTGGGGGAGACAGGGAGGCTGAGGTGTGGCCGAGGACCTGACCATCTG 10980
 GAAGTGTGAAGATCCCTTTGGGCTGTCAAGAGCCTTGGGCTTGCCATGAAGTGAAGAG 11040
 AGTGGCAGCTCTCATGAGGGTGGCGAAGGTGAAGTGAAGAGATCTACACAGAGTCCCCA 11100
 GCCTGGGCTCACCTGCACTTCTCTTCCCTCTGACCACTTTTGCGCAGCTCATCCCCG 11160

1632 (EXON 41)
 T K D V K A A A H Q M R N F L V R A S
 CAGCCAGAGATGTCAGGCGCTGCTAATCAGATGCGCACTTCTGTTTCGAGCCTCCT 11220

C R L R L E P G K E Y L I H G L D G A T
 GCCGCCCTTCGCTTGAGACCTGGGAAGGAATATTTGATCATGGGCTGGATGGGGCCACCT 11280

1676
 Y D L E G H
 ATGACCTCGAGGGACAGTGAATCATCTGGTCCCTCACTCTCTTGTCTCCCATGCTC 11340

GCCACCTAGGCTTGGCCCTCAGAGGCCAGATGCTGTGCTCTCCGTTTCACCTGCCAT 11400

1677 (EXON 42)
 P Q Y L L D S
 CCTCCCGAGCCCTGCTGACTGCCCTTTGCCCTTGCAGCCCCAGTACCTGCTGAGCTC 11460

H S W I E E H P S E R L C R S T R Q R A
 GAATAGCTGATCAGAGAGATGCCCTCTGAGCGCTGTGCCGAGCAGCCGCCAGCGGGC 11520

1722
 A C A Q L N D F L Q E Y G T Q G C Q V *
 AGCCTGTGCCAGCTCAACGACTTCTCCAGAGATATGGCACTCAAGGGTGCAGGTGTG 11580
 AGGGCTGCCCTCCACCTCCGCTGGGAGGAACTGAACCTGGGAACCATGAAGCTGGAAG 11640

Poly-A signal
 CACTGCTGTGTCCGCTTTCATGAACACAGCCTGGGACCAAGGCATATTAAAGGCTTTTG 11700

* Poly-A site
 CAGCAAGGTGTAGTGTGTCAGCGAAGTGTCAATGTGTGTTGCTAGGGCTGAGAGCAGT 11760
 GCCCCGCCCCGATGAGTTCTGGGAGGCCAGGTTGACATAACCTTAGACTCTCTGAGCC 11820
 CTGATGACCTTGGGCTGTTCAGCTCTGCTAGAACCTCCAGATGACCCGCTAGGAGTCT 11880
 AGTGCTTCACAGGACCAACCCGAGCAGAACTGGGACCCAGAGCCTGCACCCCAAGGACC 11940
 AGAGTCCATGCCAAGACCAACCTTCAGCTTCCAGGCGCTCCACTGCCCGCTGTGCCCA 12000
 GTCACACAGCCCTCAGACAGGGCTTGTGCTCAGCTGACACCTGTGACACAGCTCTTCTGC 12060
 CTCATGAGCTGTTGTCCAGCTACACCTCCCCAGCTGTCTCTGCTGTGCTGCTGGCGGTTCT 12120
 GAGGTCTGCAGATTTTAGCTGAGTTCCGGGCTGTTGAAGCCTGCTGACGCTTGGTTCTG 12180
 TTATCAGTGAAGTGAAGTGAATTTCCCGAGCTTGTGCAATCCTCAGGTCCGAGGTGTC 12240

TTCTTCCAGTTACTGGTTTCAACCAAGCCAAAATCTGACTTTGGTGTTTGTGAATCC 12300
TCTGAAGGAAGCCGCTGTTCTCCTGGGGTCTCCCTTCCCACCGGACCTGCTTAACTTTCC 12360
CCCATTTAGTGGCAGACCTGGGGTCTTCAAGAGATGACTCCGCGTCTGTCCAAGGAAGTTT 12420
GGTGAAGATCAGTTTCCGTAGAGGTCAGTGACAGTTCAAGCAAGCTGATCCAATCATTGCA 12480
CAGAAATTGAGGAATCTTTCACCTTCATGCCATGCCCTGTGCCAGGTGCCAGAGATACAGC 12540
TGCTCACTCCAG-AGTCATCGCTGGGGAGACAGATAAGAGGACGGGCAATCCCAACCTC 12600
TGTGAAGAGATGTGATGTCAGGGAGCAGGTGTGGTCCGTGGGGCATCTAACCAGGTCAGGG 12660
GCATTGCCAGGCAAGGGAAGGGAGGGCTTCTTGGAGCAAGTGGCTCCAAGTGGGGCTCTG 12720
AAGACTGAGGAAGGAGCCAGGCAAGAGAGCAAGGGGTAGATGAGGGCATCTGGGGCAAGAGGA 12780
GAATATACAAAGGCCAGAGGCCGGGGGCAAGACAGGGTACCTTTGGGGACATTGCATGT 12840
AATTGACCACATTGAGAGTTTGGATTTGGAGGTGGTGGAGAGATGGAGATGGTGAAGACA 12900
AGTAGTAGCAGCTCAGCCTTCCAGGTGCCTCTTTCCGATGAGCACTGTCTTATCCCA 12960
CGTAACCTTTGAGGAAGTTTGGGGCTTTCCCACTGTGGCAAGGGTTTCTGAGGGCTTTGCA 13020
BglII site; to 5' end of 21-OHase A
TACATGGCCCTATGGTTGCTCATCAGATC 13048

PUBLICATIONS

1. * Belt, K.T., Yu, C.Y., Carroll, M.C. and Porter, R.R. (1985). Polymorphism of Human Complement Component C4. *Immunogenetics* 21, 173-180.
2. Carroll, M.C., Belt, K.T., Palsdottir, A. and Yu, C.Y. (1985). Molecular Genetics of the Fourth Component of Human Complement and Steroid 21-Hydroxylase. *Immunological Rev.* 87, 39-60.
3. * Carroll, M.C., Palsdottir, A., Belt, K.T. and Yu, C.Y. (1985). Molecular Genetics of the Fourth Component of Human Complement. *Biochem. Soc. Symp.* 51, 29-36.
4. * Yu, C.Y., Belt, K.T., Giles, C.M., Campbell, R.D. and Porter, R.R. (1986). Structural Basis of the Polymorphism of Human Complement Components C4A and C4B: Gene Size, Reactivity and Antigenicity. *EMBO J.* 5, 2873-2881.
5. Reid, K.B.M., Campbell, R.D., Day, A.J., Dodds, A.W., Law, S.K.A., Lintin, S.J., Ripoche, J., Sim, R.D., Wu, L.C. and Yu, C.Y. (1987). Molecular Biology of the Human Complement Class III Products of the MHC; Factor I and its Cofactors. *Prog. Immunol.* In Press.
6. Yu, C.Y. and Campbell, R.D. (1987a). Sau 3A1 Polymorphism of Human Complement Component C4 at the C4d Region. *Biochem. Soc. Trans.* 5, 654-655.
7. * Yu, C.Y. and Campbell, R.D. (1987b). Definitive RFLPs to Distinguish between the Human Complement C4A/C4B Isotypes and the Major Rodgers/Chido Determinants: Application to the Study of C4 Null Alleles. *Immunogenetics* 25, 383-390.
8. * Rodrigues, N.R., Dunham, I., Yu, C.Y., Carroll, M.C., Porter, R.R. and Campbell, R.D. (1987). Molecular Characterisation of the HLA-linked Steroid 21-Hydroxylase B Gene from an Individual with Congenital Adrenal Hyperplasia. *EMBO J.* 6, 1653-1661.

*

Copies of these publications can be found in Appendix II.