Characterisation of the Genomic Region Upstream of PSG-11.

A Thesis Presented in Partial Fulfilment
of the Requirements for the degree of
Master of Science in Genetics

at

Massey University, Palmerston North
New Zealand

Terence Wayne Joe

1994

TABLE OF CONTENTS

Abstract i

Acknowledgements ii

Abbreviations iii

List of Figures, Tables and Equations iv

A: INTRODUCTION

A.1: The Placenta 1

A.2: Pregnancy-Specific β_1-Glycoprotein 1

A.3: Clinical Applications 5

A.4: The Structure and Evolution of CEA and PSG 7
A.4.1: Structure of CEA 7
A.4.2: Structure of PSG 10
A.4.3: Size of the CEA/PSG gene family 16
A.4.4: Evolution 16

A.5: Possible Biological Role(s) of the PSG 21

A.6: Objectives of this Project 23

B: METHODS AND MATERIALS

B.1: MEDIA AND SOLUTIONS 25

B.2: Bacterial Cell Cultures 26
B.2.1: Cell Genotypes 26
B.2.2: Storage of Bacterial Cell Cultures 26
B.2.3: Preparation of *E. coli* Competent Cells 27
B.2.4: Preparation of C-600 Plating Cells 27
B.2.5: Infection of *E.coli* C-600 Cells 27

B.3: DNA PREPARATION 28
B.3.1: Rapid Isolation of Double Stranded DNA 28
B.3.2: Medium Scale λ DNA Preparation 28
B.3.3: Large Scale Plasmid DNA Isolation 29
B.3.4: Isolation of Human Genomic DNA 30

B.4: STANDARD PROCEDURES 31
B.4.1: Restriction enzyme digests of DNA 31
B.4.2: Ethanol Precipitation 31
B.4.3: Agarose Gel Electrophoresis 31
B.4.4: Phenol-chloroform extraction of DNA 31

B.5: DNA CLONING PROCEDURES 32
B.5.1: Ligation of DNA Fragments 32
B.5.2: Gel Purification 32
B.5.3: Transfection and Transformation 32
B.5.4: Electrotransformation and Electroporation 33

B.6: NUCLEOTIDE SEQUENCING 34
B.6.1: Small Scale DNA Preparation for Sequencing of M13 34
B.6.2: Polyacrylamide Gel Electrophoresis 34

B.7: SOUTHERN HYBRIDISATION ANALYSIS 35
B.7.1: Vacuum Blotting 35
B.7.2: Labelling Probes 35
B.7.3: Hybridisation 37
B.7.4: Oligonucleotide Labelling and Hybridisation 37

B.8: Screening of a cDNA Library 38

B.9: Cos-mapping 39
B.9.1: End Labelling the Oligonucleotides 39
B.9.2: Oligonucleotide Hybridisation 39

C: RESULTS AND DISCUSSION OF THE hC3.11 SEQUENCE

C.1: Characterisation of hC3.11 41

C.2: The Subclone pE2.2 41
C.2.1: Titration of the Bal 31 Enzyme 43
C.2.2: The Bal 31 Strategy 45

C.3: Results for the Combined hC3.11 Sequence 49
C.3.1: The 500 bp Intron Region 51
C.3.2: Previously Unreported Splice Sites 51

C.4: Analysis of the hC3.11 Sequence 54
C.4.1: Defining the `Gene-specific' Sequence 54
C.4.2: PSG Subgroup-1-Like C-Termini54

C.5: DISCUSSION OF THE hC3.11 SEQUENCE 56
C.5.1: Is the Upstream Gene a Subgroup-3 Gene 57
C.5.1.1: Subgroup-1 57
C.5.1.2: Subgroup-2 58
C.5.1.3: Subgroup-3 58

D: RESULTS AND DISCUSSION FOR COSMID CLONES #1 AND #4.

D.1: Defining the Clones 61
D.1.1: The Clone hC3.11 61

D.2: The Probes 61
D.2.1: The `Gene-Specific' Probe 61
D.2.2: The PSG Domain Probe 62

D.3: Isolation of Clones #1 and #4 62
D.3.1: The Lorist Vector Clones 63

D.4: Results from Southern Hybridisation and Restriction Mapping 65
D.4.1: Hybridisation Analysis of Clones #1 and #4 65
D.4.1.1: Hybridisation of Cosmid Vector Arms 65
D.4.1.2: Hybridisation of an N-Terminal Probe to Clones #1 and #4 68
D.4.1.3: Hybridisation of Human Genomic DNA 70
D.4.2: Cos-Mapping Clones #1 and #4 72

D.5: Analysis of Clones #1 and #4 77
D.5.1: Cosmid Clone #1 77
D.5.2: Cosmid Clone #4 77

D.6: DISCUSSION FOR CLONES #1 AND #4 82
D.6.1: The Cos-Mapping 82
D.6.2: Clones #1 and #4 85
D.6.3: The Cosmid Vector Arms 86
D.6.4: The N-Terminal Probe 87
D.6.5: The Validity of the `Gene-specific' Probe 88

E: CONCLUSIONS 90
E.1: Future Studies 90

F: BIBLIOGRAPHY 92

G: APPENDIX 106

ABSTRACT

The genomic clone hC3.11, isolated in 1989 in our laboratory encompasses the majority of the PSG-11 gene and contains 8.5 kb of upstream intergenic sequence. Nucleotide sequence from the initial 1.5 kb of the hC3.11 clone revealed the presence of a number of unique PSG-like C-domains upstream of the PSG-11 gene. In this investigation, the sequencing of this region was completed resulting in 3.8 kb of contiguous sequence representing the area of interest. When compared to known PSG sequences the combined hC3.11 sequence was found to be similar to other PSG genes, but also contained several unique and previously unreported C-domain-like regions.

Chromosome walking techniques were used to investigate the area upstream of the PSG-11 gene. Two cosmid clones, #1 and #4, were isolated from a human genomic DNA library as potential candidates representing the a full length gene upstream of PSG-11. These were characterised by restriction enzyme mapping, cos-mapping and hybridisation analysis. Analysis of the data of these cosmid clones indicate that one of the clones represents an allelic variant of the hC3.11 region, whereas the other clone appears to contain a genomic fragment from another PSG locus.

Hybridisation analysis of the region stretching 9 kb upstream of the C-domain region of the hC3.11 clone failed to identify other PSG-related sequence. The absence of a PSG gene associated with the C-terminal domains, suggested that the hC3.11 C-domain region may be a remnant of evolutionary activity. It is proposed that the hC3.11 C-domain cluster represents a free-standing C-domain `cassette', which may be ubiquitous amongst PSG gene family members.

ACKNOWLEDGEMENTS

I would sincerely like to acknowledge and thank the many people who have supported and assisted me during the course of this degree. Firstly, many thanks to my supervisor, Dr Brian Mansfield for his much appreciated guidance, optimism and encouragement. A special thanks to Trish McLenachan, for all the enthusiasm, advice and invaluable help during this trial. Thanks also to the staff and populace of the Micro and Genetics faculty for providing an interesting setting for this drama.

To each of the members of Mansfield park who have passed through during my sentence, including Kyle, Ruth, Kay, Lester, Felix, Mike, Neville, and Sheralee, my thanks for all the help and many welcome hours of distraction.

A special mention must be made of Merie, Paul, Shalome, Joseph, Geoff, Morgan, Sheree, Delwyn and David for the laughs and the company as we endured our thesis years.

To my family for their unquestioning provision and support of my education, in particular to the memory of Amy Joe to whom this thesis is dedicated.

Finally, but foremost I thank and credit the α & θ for the Faith and Grace that began and finished this work.

ABBREVIATIONS

bp	-	Base pairs
C-terminus	-	Carboxyl terminus
CEA	-	Carcinoembryonic antigen
DNA	-	Deoxyribonucleic acid
EDTA	-	Ethylenediaminetetraacetic acid
IPTG	-	Iso-propyl β-D thioglylactosidase
kb	-	Kilobase
N-terminus	-	Amino terminus
ON-R	-	Right cosmid vector arm
ON-L	-	Left cosmid vector arm
PSG	-	Pregnancy-Specific β1-Glycoprotein
RNA	-	Ribonucleic acid
SDS	-	Sodium Dodecyl Sulphate
SSC	-	Standard Saline Citrate
U	-	Units
UV	-	Ultra-violet
X-Gal	-	5-bromo-4-chloro-3-indolyl-β-D-galactopyranosidase

LIST OF FIGURES, TABLES, AND EQUATIONS

Page

Figure 1 CEA The Prototype Domain Model 9
Figure 2 Immunoglobulin Superfamily 9
Figure 3 Pregnancy Specific Glycoprotein 13
Figure 4 Domain Organisation in PSG Proteins 13
Figure 5 The Three PSG Subgroups 15
Figure 6 Evolution of the CEA Gene Family 20
Figure 7 Subclones of Cosmid hC3.11 42
Figure 8 Titration of the Bal 31 Enzyme 44
Figure 9 The Bal 31 Deletion Clones 46
Figure 10 The Combined hC3.11 Sequence 47-48
Figure 11 The Splice Consensus Sequence 49
Figure 12 The Subgroup-1 Arrangement 60
Figure 13 The Subgroup-2 Arrangement 60
Figure 14 The Subgroup-3 Arrangement 60
Figure 15 Cosmid Clones 64
Figure 16 Patterns of Hybridisation Clone #1 66
Figure 17 Patterns of hybridisation Clone #4 67
Figure 18 N-terminal Oligonucleotide Hybridisation 69
Figure 19 Hybridisation of Genomic DNA 71
Figure 20 A Typical Cos-mapping Gel 73
Figure 21 Maps for clones #1 and #4. 74
Figure 22 Standardised Maps for Clone #1 75
Figure 23 Standardised Maps for Clone #4 76
Figure 24 Three Scenarios for Clone #4 79
Figure 25 Products of Cos-mapping Digests 84

Page

Equation 1 Efficiency of labelling (% Incorporation)....37
Equation 2 Calculation of Specific Activity.... 37
Equation 3 Stringency, Calculation of T_m.... 39
Equation 4 Calculation of Rate of Activity for Bal 31 enzyme.... 43
Equation 5 Fragment Length Standardisation Equation....72

Table 1 Placental Protein Products.... 4
Table 2 PSG C-terminal Domain Variants.... 14
Table 3 The CEA Gene Family.... 18
Table 4 Standardised Nomenclature.... 19

INTRODUCTION

A.1: The Placenta

One of the greatest immunological challenges occurring naturally in humans is presented by development of the feto-placental unit during pregnancy. The successful allogeneic grafting of the fetal to maternal tissue takes place, despite the existence of a complex maternal immune system, which would be expected to defeat fetal implantation. A major factor in the survival of the placental allograft is the specific lack of rejection by the mother toward the developing foetus {124}. Although the basis of this immunological tolerance is not completely understood, it is likely that the relationship between the mother and the developing foetus involves basic humoral responses.

Experiments have shown that serum immunoglobulin concentrations increase during pregnancy, whereas cell-mediated responses decrease, suggesting that humoral response plays a predominant role in this tolerance. However, the multiplicity of molecules involved, some appearing at different times and in varying amounts, complicate the identification of the specific immunoglobulin subclasses involved {85}.

The human placenta is an organ comprised of tissues of two different genotypes, providing an `interface' for exchange of gases and nutrients between the mother and the developing foetus, while still preserving the individuality of both systems {85,124}.

Not only does the placenta temporarily serve as a fetal lung, kidney, liver and intestine, but it also acts as an active exocrine and endocrine gland. An array of complex endocrine functions are initiated and completed by the placenta, completely taking over functions of the maternal ovary and pituitary gland. A wide variety of hormones, enzymes, growth factors and other molecules essential for the survival and development of the human foetus are produced by the placenta {85,119,120,121}.

A.2: Pregnancy Specific Glycoprotein

Many pregnancy-specific proteins have been reported from gel electrophoretic and immunological studies of placental extracts and maternal sera. These are summarised in Table 1.

The first in this group to be identified was the Pregnancy-specific β_1-glycoprotein (PSG), originally known as pregnancy associated plasma protein C (PAPP-C).

In 1970, Tatarinov and Masyukevich isolated PAPP-C as a new protein from the serum of pregnant women {1}. This was soon discovered to be immunologically identical to the placental protein SP-1 (Schwangershafts protein-1) isolated by Bohn in 1971 {2}.

Initial studies focused on the development of sensitive assays for the protein. It was found that during human pregnancy, the protein SP-1 was present at the highest levels of any placenta-specific protein in maternal serum (350 µg/ml) {3}.

Human PSG are primarily synthesized in the placental syncytiotrophoblast cells during pregnancy, and are subsequently secreted into the maternal serum {1,2,4,5}. The secretory nature of the PSG has been demonstrated in vitro both in primary culture of trophoblasts {6}, and by transfecting cloned PSG cDNA into cultured mammalian cells {7,8,9,10}.

The PSG protein is first detectable in human serum 7 days post conception {11}. The levels increasing exponentially during pregnancy with a doubling time of 2-3 days and a half-life of 30h, to reach term concentrations of 200-400 µg/ml maternal serum {1,12}.

As with the majority of placental products, expression of the PSG protein is not limited exclusively to the placenta {99}. Low levels of PSG protein have been detected in non-placental tissue as well as various cell lines {2,5}. Studies have demonstrated PSG gene expression in non-placental tissues {17}, human fibroblasts {35,36,37,38} and malignant tumor cells {39,40}. Isolation of PSG clones from cDNA libraries created from testis {24}, fetal liver {8,26}, salivary gland {27}, intestine {27,41} tissues, HeLa {16} and myeloid cell lines {22,30} provided further evidence showing the expression of PSG in non-placental tissue.

Biochemical studies demonstrated that the human PSG gene products are actually a heterogeneous group of proteins consisting of at least 3 distinct placental protein species with molecular weights of 72, 64, and 54 kDa {11,14,15,16}. These immunologically similar proteins were found to have carbohydrate contents ranging from 28%-32% {14}.

To study their function(s) and to develop specific reagents for the individual PSG, it was considered important to identify all the PSG genes.

Despite the previous biochemical analysis of several PSG species, the complexity of the family was not fully appreciated until the PSG genes were cloned {17,18}. Several PSG cDNA clones were independently isolated and characterised by Watanabe and Chou {14}, Streydio et al. {15} , Rooney et al. {11}, Chan et al. {16} in 1988. Additional PSG cDNA's {7,9,10,19-34} and genes {20,22,25,29,34} were subsequently reported. Examination of these genes revealed the conserved nature of the PSG family, capable of producing an array of highly related, yet unique gene products.

PROTEIN	ABBREVIATION	ANALOGUE IN NON-PREGNANT ADULT	MOL.Wt. (kDa)
Human chorionic Gonadotrophin	hCG	Luteinising hormone	45-50
Human placental Lactogen	hCS	Prolactin, growth hormone	21-23
Human chorionic Thyrotrophin	hCT	Thyroid stimulating hormone	45
Human chorionic corticotrophin	hCCT	Adenocorticotrophic steroid	5
Human chorionic gonadotrophin releasing hormone	hC-LRH	Gonadotrophin releasing hormone	1
Scwangerschafts-spezifisches β1 glycoprotein	SP-1	Unknown	90-110
Pregnancy specific β1 glycoprotein	PSβG		
Trophoblast specific β1 globulin	TBG		
Pregnancy associated plasma protein C	PAPP-C		
Pregnancy associated plasma protein A	PAPP-A	Unknown	750
Pregnancy associated plasma protein B	PAPP-B	Unknown	1000
Heat Stable Alkaline Phosphotase	HSAP	Alkaline Phosphotase	
Cysteine aminopeptidase (oxytocinase)	CAP	Aminopeptidases	
Diamine oxidase (Histaminase)	DO	Histaminase	190
Placental protein 5	PP 5	Unknown	42

Table 1: A list of some of the many protein products produced by the human placenta. Adapted from Klopper *et al.* (128).

A.3: CLINICAL APPLICATIONS

Soon after the fundamental studies of Bohn {46,47} several clinical groups explored the possibility that the PSG could be of prime importance in the evaluation of pregnancies. This interest centred on both early and late pregnancy, and also on the production of PSG by tumours.

Bohn et al. demonstrated that PSG were essential for the maintainence of human pregnancy by showing that antibodies to PSG induced abortion in primates {48}. Another study using non-pregnant monkeys actively immunised with PSG, resulted in a loss of fertility with subsequent pregnancies often ending in abortion {49}. Moreover, a correlation between low maternal serum PSG levels during pregnancy and threatened abortion, was observed emphasising the importance of PSG in the maintainence of healthy primate pregnancy {49}.

The development of sensitive assays for PSG in maternal serum {50,51,52}, allowed several pregnancy related complications to be predicted, when used in conjunction with other tests and indicators.

For example, low levels of PSG in maternal serum during pregnancy can be indicative of ectopic pregnancy {57,58}, and when used in conjunction with ultrasound and/or a human placental lactogen test, threatened abortion can be predicted with approximately 97% accuracy {59,60,61}.

Such conditions as foetal intrauterine growth retardation and intrauterine foetal death are associated with low levels of PSG and are diagnosed in conjunction with ultrasound scans.{62}. The routine method of diagnosing Meckel's syndrome is by measuring the high concentrations of PSG in the amniotic fluid associated with this condition {63}.

One of the placental gene products often used in diagnosis of pregnancy related conditions, in conjunction with PSG, is the well characterised hormone, human chorionic gonadotrophin (hCG). This hormone, hCG, first detected in the serum and urine of pregnant women by Aschheim and Zondek in 1927 {53}, was demonstrated to be produced by the syncytiotrophoblast {54}, and was subsequently established as the most reliable marker of a viable trophoblast.

The pregnancy test of choice involves the measurement of human chorionic gonadotropin (hCG). PSG levels are used as an adjunct to the hCG test, and are also used to detect Gonadotrophin induced pregnancies {55}.

Babies afflicted with Down syndrome can be predicted with 72%-78% confidence when the concentration of PSG is measured in conjunction with human chorionic gonadotrophin (hCG), and α-fetoprotein and assessed along with maternal age {56}.

Human tumour cells have been found to produce immunoreactive PSG {64}, therefore, PSG have been used as a marker in the diagnosis of certain cancers and tumours. Since high levels of PSG have been associated with choriocarcinoma, hydatidiform mole and gestational trophoblastic disease, the concentration of PSG has been used both as an indicator, and as a prognosis index in the treatment of these conditions {5}. Searle et al. (1978) {66}, suggested that serum PSG concentrations alone are not of great value in the detection and monitoring of carcinoma in the breast, large bowel and ovary, since the increase in the concentration of PSG associated with these conditions, does not correlate with the extent of the disease. Therefore, as with pregnancy related conditions, the clinical measurement of PSG is usually performed in conjunction with other tests in the diagnosis of these tumours.

Measurement of human chorionic gonadotrophin (hCG) in plasma or urine is widely used in the diagnosis and management of trophoblastic tumours. Since the concentration of hCG is related to the total cell mass of the tumour cells, the rate of cell growth and regression of the tumour cell population can be predicted {65}.

The ratio of the two placental proteins hCG and PSG forms a prognosis index for hydatidiform mole and gestational trophoblastic disease. A value less than 5 indicates a 73% chance of persistent disease, whereas a value greater than or equal to five, indicates a 74% chance of remission {67}. In the treatment of breast cancer patients, the absence of PSG suggests an improved prognosis, while presence of PSG in breast cancer patients estimates a 40%-85% chance of mortality in less than 4 years {68,69,70}.

In the management of trophoblastic tumours, the measurement of hCG concentration is the most useful measurement in these patients. Only in isolated cases in which PSG persists after hCG has become undetectable, does the measurement of PSG become valuable {100}.

The determination of PSG concentration in the serum of patients with non-trophoblastic tumours such as carcinoma of the breast, intestine or ovary, does not give practical information on the extent and progression of the disease. However, detection of PSG in carcinoma tissue itself may have prognostic significance. Since PSG is said to have immunosuppressive properties {71,72}, it

seems likely that the production of this protein by malignant tumours might be a means by which the tumour escapes immunological recognition and continues to grow.

Therefore, there are possible practical implications in investigating the involvement of the PSG in circumventing the human immune system, to allow the design of more effective treatments for tumourous conditions.

Investigation into the nature of phosphorylation in PSG proteins could provide insight into the role of PSG in pregnancy and diseases. Phosphorylation of tyrosine residues in proteins, have been shown to play a major role in the control of cell growth and differentiation {101}. Therefore, similar events in the PSG could conceivably trigger a cascade of events involved in implantation and trophoblastic invasion.

A.4: THE STRUCTURE AND EVOLUTION OF CEA AND PSG

A.4.1: Structure of CEA

The human PSG are encoded by multiple, linked genes located on chromosome 19, q13.1-13.3 overlapping the region containing the closely related CEA gene subgroup {74,75,77}. Fluorescence in situ hybridisation to metaphase chromosomes localised the PSG subgroup telomeric to the CEA subgroup. Finer mapping suggests that most of the genes are contained within 800 kb of sequence flanked by SacII restriction sites {77}. In total, the CEA/PSG gene family region is estimated to span 1.1 to 1.2 Mb {76}.

Based on sequence comparisons, the PSG have been classified as a subgroup of the CEA family, for which carcinoembryonic antigen is the prototype.

Carcinoembryonic antigen (CEA) was found to be present in colonic tumors and foetal gut tissue by Gold and Freedman in 1965 {104}. It was initially thought to be absent in normal adult intestine, however, later studies revealed the presence of CEA in several normal tissues including human colon {105-108}. Despite the lack of tumor specificity, CEA is one of the most widely used human tumor markers for assessing the recurrence of colorectal, breast and lung cancers. The serum concentration of CEA represents an important parameter in the post-operative surveillance of cancer patients {109,110}

The CEA protein is a highly glycosylated molecule with a molecular weight of 180,000 daltons. Glycosylation inhibition studies show the protein to consist of a single polypeptide chain with an apparent molecular weight of approximately 80,000 daltons {111}. Amino acid sequence deduced from the nucleotide sequence of the CEA cDNA, shows that CEA is synthesised as a precursor of 702 amino acids. The leader peptide (34 amino acids) is followed by the mature CEA peptide (668 amino acids).

Due to the presence of three internal repeats, the peptide can be divided into a number of structural domains. A schematic diagram of the CEA domain arrangement is shown in Figure 1. The three repeat domains of 178 amino acids each reveal an exceptionally high degree of sequence similarity, having between 67% and 73% of their amino acids identical. Allowing for conserved changes/substitutions the degree of similarity is even higher. Each domain contains four cysteine residues at precisely the same positions. These CEA repeat domains are relatively long, and each can be further sub-divided into two subdomains (A,B) of approximately equal size. The amino acid sequence from these subdomains display similarity to each other, as well as to the N-terminal domain {123}. The degree of conservation at the nucleotide level is also very high (80%-83% identity). Other proteins with internal repeats have been reported in the literature, but the internal degree of similarity of the repeating domains of CEA is the highest reported so far {112}.

Analyses at the genomic level for members of the CEA family indicated a precise correlation between the exons and the A and B sub-domains {116,117,122}. A domain model was subsequently proposed for CEA, which assumes that the conserved neighbouring cysteine residues present in the repeat domains form disulphide bonds, creating a looped secondary structure characteristic of the immunoglobulin family {113,116,123}. This is shown in Figure 2, demonstrating the strong similarity in secondary structure amongst members of the immunoglobulin superfamily {84}.

The C-terminal domain of CEA consists of 27 amino-acids and is strongly hydrophobic. This provides a potential insertion region for the CEA protein into the plasma membrane and is of an appropriate length to span the lipid bilayer. Therefore, this provides a possible means to anchor CEA to the cell surface membrane {124}.

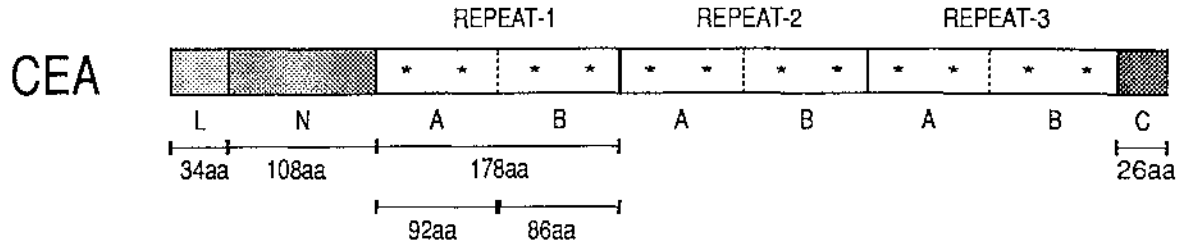

FIGURE 1: CEA THE PROTOTYPE DOMAIN MODEL.

The CEA protein domain arrangement based on deduced amino acid sequence. Domain sizes are indicated below the domain blocks. Invariant cysteine residues are marked (*). Adapted from Thompson et al. {25}.

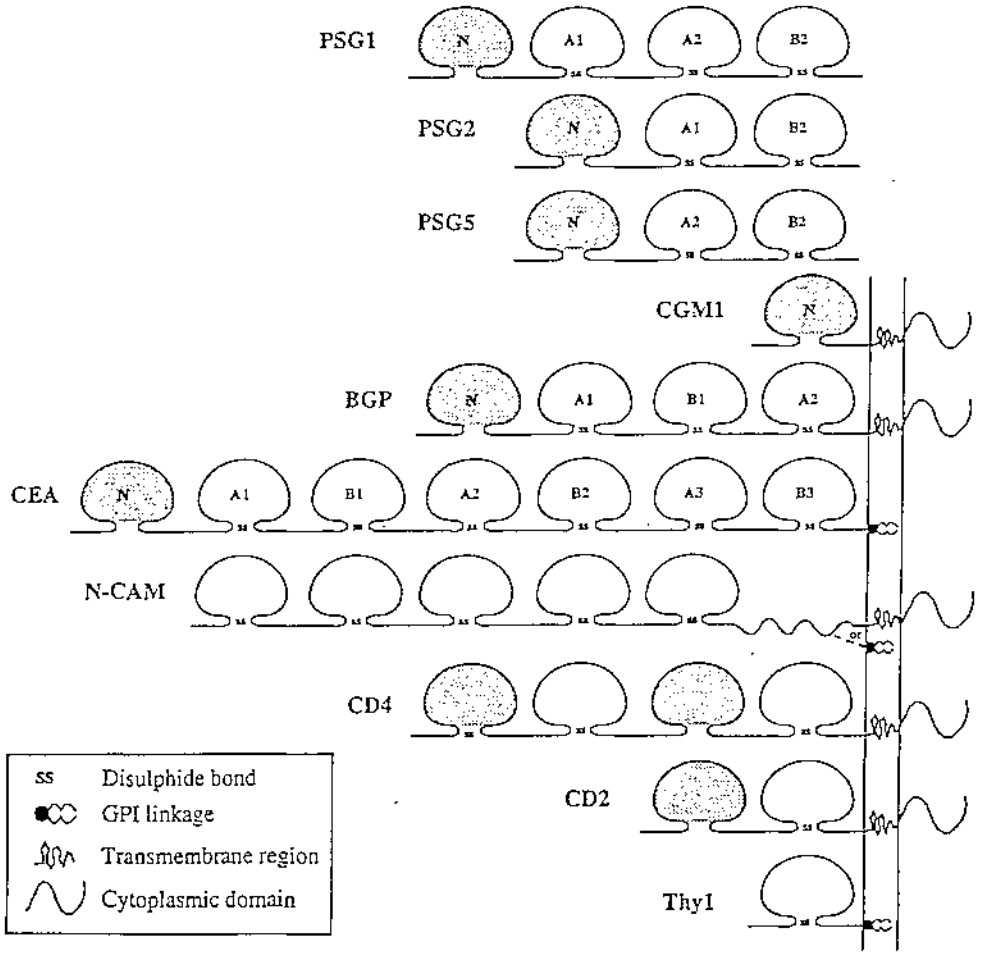

FIGURE 2: MEMBERS OF THE IMMUNOGLOBULIN SUPERFAMILY

A schematic representation of some members of the immunoglobulin family. Filled circles represent variable region (V)-like domains, other circles closed with (ss) represent constant domains. Adapted from Khan et al. {81}.