# Characterisation of the 3' Region of the PSG11 Gene

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.

**Patricia Ann McLenachan**

**1995**

# ABSTRACT

The pregnancy-specific beta-1 glycoproteins (PSG) are a family of abundant proteins essential to pregnancy that are encoded by 11 genes located on chromosome 19q 13.1-13.3. The genes can be divided into three subgroups based on the organisation of their 3' coding regions. In 1989, our group isolated cosmid hC3.11, which contained most of the PSG11 gene, but which did not include the 3' coding region. This thesis reports subsequent work to characterise two further cosmids which span the PSG11 locus and which do include the 3' coding region. These cosmids were mapped and partially sequenced.

Three exons encoding potential alternative C-terminal domains were identified: Cw, Cr and Cs. The Cs domain lies 4.6kb from the end of the B2 domain. This is the first report of genomic sequence for this particular domain and for a functional PSG subgroup 3 gene.

Downstream from this exon are sequences homologous to the C-termini of subgroup 1 PSG genes. This finding suggests that subgroup 1, 2 and 3 genes are related via insertion/deletion events.

Data from seven PSG genes from all three subgroups and from four different regions were used to construct evolutionary trees. Variability patterns in the data were examined and these showed that the mechanism of sequence evolution for the N-terminal domain, the A1 domain, and to a certain extent, the B2 domain was not neutral. Sequences from these regions were shown to be unsuitable for determining historical relationships between PSG genes. In contrast, the data from the C-terminal region showed a better fit with the assumptions of sequence evolution (e.g. all changes are independent and identically distributed) required by currently implemented analysis methods. Evolutionary tree reconstruction using this region gave a phylogeny that was consistent with one based on the genomic organisation of the genes.

# ACKNOWLEDGEMENTS

This thesis has been a long time in the making, consequently, the number of people who have provided inspiration, encouragement and practical support has been large.

Many thanks must first go to my supervisor, Dr.Brian Mansfield, for whom I have worked for the past eight years. During this time I have picked up many tricks and techniques but have, above all, gained an increasing confidence in myself due to Brian's support and encouragement. To all the people who, over the last eight years, have passed through the now disbanded Mansfield Park, a special thanks for sharing the highs and the lows that come with research in the field of molecular biology. To Kay Rutherfurd, Neville Honey, Max Scott and the present-day members of the Cake Club - a special thanks and please let me know when you celebrate the next good result.

My Head of Department, Professor Tim Brown, has been steadfastly encouraging and has provided practical support. Also, I have been fortunate to have been able to discuss ideas along the way with Professor David Penny and Dr. Robert Hickson.

Special thanks go to the past and present members of Scott Base, especially Carolyn Young, for working and playing hard.

Eric and Betty Terzaghi have been a constant source of encouragement, inspiration and practical support over the years. Heart-felt thanks to you, we miss you.

The writing of this thesis has resulted in the complete disruption of my household for an extended period of time- to my children, Sam, Amy and our latest addition Toby - thanks for your help and foreberance. Special thanks too, to my mother Rita, for looking after all of us so well, during the final throes.

To my very special friends Rick, Kathy and Megan, thanks for looking after the boy and for Friday night take-outs and videos. Hope we can do it all again.

To Amy and Adalie for lunches at Bellas, baby therapy and your special friendship, many thanks.

To Maren, Brian ,Karen and David, many thanks for fun, friendship and lots of encouragement.

And last of all, to my husband Peter - for discussions and ideas, for your hard work, your constructive criticism and encouragment, for minding Toby and for taking me to Bath - what can I say- thank you, thank you, thank you, I owe you bigtime.

## ABBREVIATIONS

| | |
|---|---|
| aa | amino acid |
| ATP | adenosine tri-phosphate |
| B-ME | beta-mercaptoethanol |
| BSA | bovine serum albumin |
| CGM | **c**arcinoembryonic antigen **g**ene family **m**ember |
| CIAP | calf intestinal alkaline phosphatase |
| DMF | dimethylformamide |
| DNA | deoxyribonucleic acid |
| DNase | Deoxyribonuclease |
| DTT | Dithiothreitol |
| *E.coli* | *Escherischia coli* |
| EDTA | ethylene-diamine tetraacetic acid |
| EtBr | ethidium bromide |
| IPTG | Isopropyl B-D-thioglylactoside |
| KOAc | potassium acetate |
| NaOAc | sodium acetate |
| PCR | Polymerase Chain Reaction |
| pfu | plaque forming unit |
| PSG | pregnancy specific β-1 glycoprotein |
| PVP | polyvinylpyrididine |
| RNase | Ribonuclease |
| SDS | sodium dodecyl sulphate |
| SSC | Standard Saline Citrate |
| Tris-HCl | Tris hydroxy-methyl aminomethane, made to the appropriate pH with concentrated HCl |
| TE | Tris-EDTA Buffer |
| TBB | Tris Borate Buffer:. |
| TAE | Tris Acetate EDTA Buffer: |
| UTR | untranslated region |
| UV | ultra violet |
| vol | volume |
| X-gal | 5-bromo-4-chloro-3-indolyl-B-D galactopyranoside |

## LIST OF FIGURES.

## List of Tables.

# CHAPTER 1. INTRODUCTION

## 1.1 The Placenta

The placenta is the organ primarily responsible for the interchange of substances between the developing foetus and the mother. As such, it performs a number of functions such as gaseous exchange, the supply of nutrients, the removal of waste products, the production and exchange of proteins and hormones as well as providing some immunological and physical protection as a barrier. The placenta is therefore a unique organ with several distinctive characteristics. It has a dual origin being derived from both maternal and foetal tissues. It has extensive contact between foetal chorionic villi and the maternal blood system. Because of its function it has both a limited life span and an extracoporeal position with respect to the foetus (Hamilton *et al.*, 1972).

Although much of the interchange of materials between the foetus and the mother is achieved by simple diffusion, the placenta cannot be considered to be an inert barrier. There is evidence that many substances, including nutrients are actively transported against a concentration gradient. Pinocytosis is thought to be involved in the transport of some immunologically important macromolecules from the maternal blood system to the foetus (Hamilton *et al.*, 1972). In a recent theoretical paper (Guilbert *et al.*, 1993), the placenta is likened to a completely syncitialised macrophage which "surrounds and protects the antigenically foreign conceptus in a manner analogous to the formation of cyst-like structures by giant cells (syncitialised macrophages) that envelop potentially harmful pathological agents". Indeed, it appears placental tissue and macrophages share many characteristics such as phagocytosis, syncitialisation, invasiveness and the expression of proteins such as CD4, CD14, IgG receptor (FcR), colony stimulating factor-1 (CSF-1), granulocyte macrophage-CSF (GM-CSF), interleukins 1 and 6 (IL-1, IL-6), tumour necrosis factor (TNF), transforming growth factor (TGF), platelet derived growth factor (PDGF) and receptors for these cytokines.

### *The immunological relationship between mother and foetus*

The placenta and developing foetus are not antigenically inert either. Why this unit is maintained, without eliciting an immunological graft rejection response, is a question for which there is as yet no satisfactory explanation. The unique immunological relationship between the mother and the developing foeto-placental unit is extremely complex. Two hypotheses proposed for explaining the specific non-reactivity of the maternal immune system to the conceptus are immuno-tolerance and immunoenhancement.

When the immune system is exposed to either very high or very low doses of antigen, it may fail to respond to the antigen; this is known as immuno- tolerance. It is possible that during gestation, small numbers of foetal cells sloe off into the maternal blood system, providing a low dose of antigen. Although the two systems are quite separate, there is some evidence that foetal erythrocytes, lymphocytes and syncitial sprouts mix with the maternal blood. Erythrocytes do not possess transplantation antigens and the antigenic status of the sprouts is unknown. Lymphocytes however, probably do possess transplantation antigens and may contribute to "low zone" immuno-tolerance (Hamilton *et al.*, 1972).

Immunological enhancement is "the specific frustration of both the antigenic stimulus and the hosts cellular immune response, mediated by humoral isoantibody" (Beer and Billingham, 1976.). The mechanism of enhancement is not fully understood but it seems to involve the masking of antigen by antibody which inhibits attack by cytotoxic T cells and/or helper T cells and/or blocking of receptors on cytotoxic T cells by antigen shed from the target cell (Roitt, 1980). Some kind of central inhibitory action by antibodies or antibody/antigen complexes, to impair the development of the host cell-mediated immune response is also thought to be involved (Beer and Billingham, 1976).

Some aspects of the unique immunological relationship between the mother and the developing foetus are partially understood:

The maternal immune system appears to be "cognisant" of the antigenically foreign conceptus. Indeed, recognition of genetic disparity seems to be essential for successful implantation and placental expansion (Guilbert *et al.*, 1993, Beer and Billingham, 1976). Also it appears that the uterus is able to express a normal immune response in other circumstances - that is - it does not appear to be an immunologically privileged site (Hamilton *et al.*, 1972).

There is some evidence from mouse and human studies, that pregnant individuals are biased in their immune response, toward humoral (antibody) responses and away from cell-mediated immunity (Guilbert *et al.*, 1993, Mosmann and Moore, 1991). $T_H1$ and $T_H2$ lymphocytes are at least partially responsible for cell-mediated and antibody responses respectively and produce a number of cytokines that cross-regulate an equilibrium between the two subsets of cells in non-pregnant individuals. Interleukin-10 (IL-10) is one such cytokine. It is produced by $T_H2$ cells and inhibits the synthesis of cytokines by $T_H1$ cells. IL-10 has been found to be expressed constitutively at the foeto-maternal interface during all three trimesters of pregnancy (Guilbert *et al.*, 1993). Other cytokines such as TGF-B, have been similarly detected. As well as functioning as growth stimulators, with a role in placental expansion, they may also help to establish the switch toward antibody responses during pregnancy. Interactions at the level of cytokine regulation are however, extremely complex and require further investigation to uncover all the pathways involved.

### 1.2 The PSG - A Major Class of Placental Proteins

One of the most abundant classes of proteins produced by the placenta are the pregnancy-specific β-1 glycoproteins (PSG). The PSG are a heterogeneous family of proteins whose levels increase in an exponential manner during pregnancy, reaching final concentrations of 200-400μg/ml in maternal serum (Oikawa *et al.*, 1989, Lin *et al.*, 1974). High levels of PSG during pregnancy correlate with stable pregnancy while decreased levels are associated with threatened miscarriage (Tamsen *et al.*, 1983).

The PSG are classified as a subgroup of the carcinoembryonic antigen (CEA) gene family on the basis of sequence identity. Both PSG and CEA proteins are predicted to contain the looped domains characteristic of members of the immunoglobulin superfamily (Paxton *et al.* 1987, Zheng *et al.*, 1990 and FIG.1). The observation of high levels during stable pregnancies, suggests the PSG may have an essential role during pregnancy, though this has yet to be

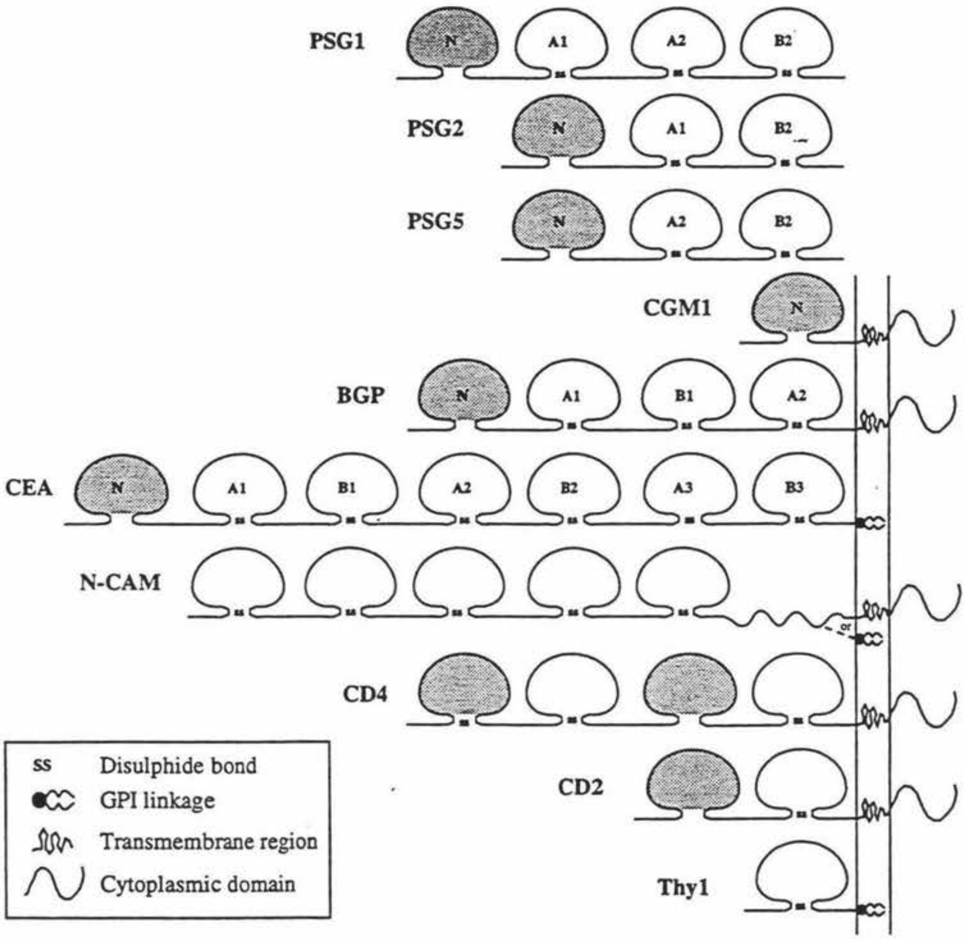


**FIG. 1. Schematic representation of some members of the immunoglobulin superfamily (from Khan *et al.*, 1992).**

Filled circles represent variable region (V) - like domains. Circles closed with the letters ss represent C2-type constant domains. Homologous domains are indicated with the letters N, A1, A2, B1, B2.

determined. Detailed characterisation of the PSG genes will provide a basis for future research into a biological role for PSG.

### *A brief historical perspective*

In 1970, Tatarinov and Masyukevich isolated a new protein from human pregnant serum which they called PAPP-C (Tatarinov and Masyukevich, 1970). A year later, Bohn isolated a protein from human placenta named Schwangerschafts protein-1 (SP1), which proved to be immunologically identical to PAPP-C (Bohn, 1971). The protein was later found to be a collection of glycoproteins with molecular masses between 70kda and 32kda and a carbohydrate content of 28-32% (Watanabe and Chou, 1988a) and were subsequently called pregnancy-specific β-1 glycoproteins. Initially PSG were thought to be specific to the placenta but later studies detected the proteins in foetal liver cells (Zimmerman *et al.*, 1989), salivary gland and colon (Zoubir *et al.*, 1990), testes (Borjigin *et al.*, 1990), in various tumour cell lines, as well as in the sera of patients with hydatidiform mole, invasive mole and choriocarcinoma (Chou and Plouzek, 1992). However, the placenta remains the major site of production of PSG.

Early experimental work using antibodies to the protein showed that human PSG were a heterogeneous group of proteins, which were thought to be related via post-translational modifications. The extent of this heterogeneity was not appreciated however until molecular studies were undertaken and it became apparent that there were several PSG genes. By the late 1980s, a number of cDNA clones had been independently isolated (Watanabe and Chou, 1988a and b, Rooney *et al.*, 1988, Chan *et al.*, 1988a, Streydio *et al.*, 1988, Khan and Hammarstrom, 1989, McLenachan and Mansfield, 1989, Zimmerman *et al.*, 1989). Subsequently, genomic clones were isolated and the genomic organisation of some PSG established (Oikawa *et al.*,1988, Oikawa *et al.*, 1989a). The genes were mapped as a cluster on chromosome 19 (Streydio *et al.*, 1990, Thompson *et al.*, 1990) and estimates of the size of the family based on the number of 5' exons detectable in the genome, suggested that there were at least 9 different genes. Many of these had alternative splice and/or polyadenylation variants, giving rise to more than thirty different transcripts.

## Table 1. CLASSIFICATION OF THE HUMAN CEA/PSG GENE FAMILY.

(from: Chou and Plouzek, 1992, Barnett and Zimmerman, 1990)

**CEA subfamily.**

| current name of gene or mRNA | old names of genes or mRNA |
|---|---|
| CEA | CEA |
| CEAa | CEA |
| CEAb | CEA |
| NCA | NCA |
| BGPa | BGPI, TM-1 |
| BGPb | TM-2 CEA |
| BGPc | TM-3 CEA |
| BGPd | TM-4 CEA |
| BGPe | 4-22 |
| BGPf | 4-13 |
| BGPg | W211 |
| BGPh | W233 |
| BGPi | W239 |
| CGM1 | hsCGM1 |
| CGM1a | CGM1a, W264 |
| CGM1b | W282 |
| CGM1c | CGM1c |
| CGM2 | hsCGM2 |
| CGM6 | hsCGM6, GN-1, M6, NCA-W272 |
| CGM7 | W236 |
| CGM8 | CGM8 |

**PSG subfamily.**

| current name of gene or mRNA | old names of gene or mRNA |
|---|---|
| PSG1 | PSβG |
| PSG1a | PSG93, PSβG-D, hPSP11, FL-NCA-2, hPS3, PSG1a, PSβG81 |
| PSG1b | PSG16 |
| PSG1c | PSβG-C |
| PSG1d | FL-NCA-1, PSG1d, SG9 |
| PSG1e | PSβG-Ci, PSG95 |
| PSG1-IIa | PSβGD |
| PSG1-I | PSG1-I |
| PSG2n | PSβG-E, SG8, hPS184 |
| PSG3m | pSP-i, hC17, PS35, hTS16, PSβGA, SG5, hPS173 |
| PSG4 | PSG4, hsCGM4, hHSP2, FL17 |
| PSG4a | PSG4, hPS133, PSG9 |
| PSG5 | PSG5 |
| PSG5-In | FL-NCA-3, hPS176 |
| PSG5-Im | PSβG-HL (clone 22) |
| PSG6 | hsCGM3, FL26, PSGGB |
| PSG6r | PSG6 |
| PSG6s | hPS12, PSG10, hPS89 |
| PSG7 | PSG7, PSGGA |
| PSG8 | CGM35, PSG8 |
| PSG8a | hTS1 |
| PSG11s | PS34, PSG-G, PSβG B, PSG7 |
| PSG11-Iw | hPS2, hPS91 |
| PSG14 | PSG14 |
| PSG15 | PSG15 |
| PSG16a | PSG9 |

In 1989, following a combined CEA/PSG workshop in Freiberg, Germany, a standard nomenclature was proposed by Barnett and Zimmerman (1990) and adopted world-wide. The current gene and transcript designations are presented in Table 1.

The PSG genes are numbered 1 through 16 , e.g. PSG1. The C-terminal splice variants are indicated with a lower case letter, eg.PSG1a, while central domain splice variants are denoted by the roman numerals e.g. PSG1-IIa (Table 1 and FIG.2).

Most recent studies of the PSG include the fine mapping of the PSG and CEA genes to determine their relative positions and orientations on chromosome 19q13.1-3 (Brandriff *et al.*, 1992, Thompson *et al.*, 1992, Trask *et al.*, 1993, Tynan *et al.*, 1992, Olsen *et al.*, 1994, Teglund and Hammarstrom, 1994), promoter analysis (Lei *et al.*, 1992, Chou and Plouzek, 1992) and initial studies on differential expression of specific PSG during pregnancy ( Streydio and Vassart, 1990, Chamberlin *et al.*, 1994). The evolutionary history of the PSG genes has been investigated and an evolutionary tree based on N-terminal nucleotide sequences proposed by Khan *et al.* (1992). A possible biological role for the PSG as haematopoetic growth factors has recently been investigated by Wu *et al.* (1993).

### *Primary structure of the PSG*

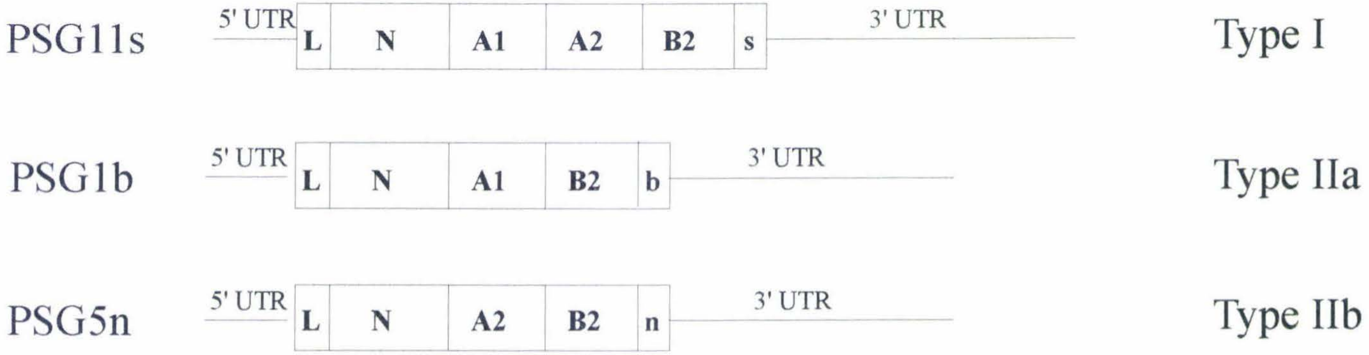
To date no direct protein sequences have been determined for the PSG. There are also no crystal structure determinations.

The primary structure of PSG, predicted from cDNA sequence, is as follows:
a short leader peptide, (L), typically 34 amino acids long, is followed by an N-terminal domain (N) of 109 amino acids (except for PSG6 which has 108), one or two similar central domains (A) of 93 amino acids, a single related central domain (B) of 86 amino acids and a C-terminus (C) of varying length (3-81 amino acids -see Table 2). The PSG proteins can be divided into two types depending on the number of central domains occurring in the protein. Type I proteins contain 3 central domains, always A1, A2 and B2. Type II proteins contain only 2 central domains, either A1, B2 (type IIa) or A2, B2 (type IIb) (FIG. 2, Chou and Plouzek, 1992).

**FIG. 2. Structure and organisation of the PSG cDNA and genes**

The domain structure of the PSG cDNA is predicted from amino acid sequence. There is a direct correspondence between the domains and the exons.All PSG genes contain six exons. The first (L), encodes part of the leader peptide. The second (L/N), encodes the rest of the leader peptide and a complete Ig-V-like, N-terminal domain (N). The central exons (A1, B1, A2, B2) encode the IgC2-constant-type domains. The B1 exon is non-functional in all PSG characterised to date. The PSG transcripts can vary in the number of central domains they possess (Types I, IIa and IIb). Also, each PSG gene can produce multiple transcripts via alternative splicing of the C-terminal exons (Ca, b, c, d, m/n, w, r and s). The PSG genes can be divided into subgroups 1, 2 and 3 on the basis of the organisation of their 3' regions.The dotted line indicates unavailable sequence data.

## SOME cDNA OF THE PSG SUBFAMILY

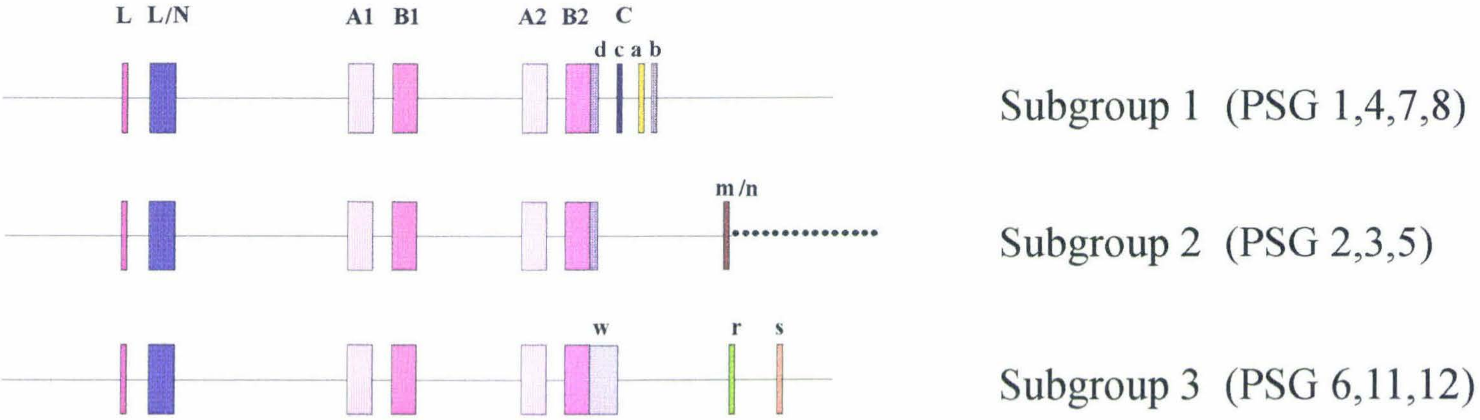


PSG11s 5' UTR L N A1 A2 B2 s 3' UTR Type I

PSG1b 5' UTR L N A1 B2 b 3' UTR Type IIa

PSG5n 5' UTR L N A2 B2 n 3' UTR Type IIb

## GENOMIC ORGANISATION OF PSG GENES



L L/N A1 B1 A2 B2 C

d c a b

Subgroup 1 (PSG 1,4,7,8)

m/n

Subgroup 2 (PSG 2,3,5)

w r s

Subgroup 3 (PSG 6,11,12)

There is extensive similarity among members of the PSG family. The N-terminal domains of different PSG share around 90% and 96% similarity at the amino acid and nucleotide levels respectively, the A domains share 91% and 96% similarity respectively and the B domains share 86% and 94% similarity respectively. In addition, the N, A and B domains are similar to each other.

The C-terminal domains initially appeared to be the most variable part of the PSG . However as the genomic organisation of the different subgroups became known, it was found that comparable regions at the C terminus also share considerable nucleotide and amino acid similarity, varying between 63% (PSG 6 and 8), and 98% (PSG 2 and 5). A comparison of the C-terminal domains of reported PSG transcripts is presented in Table 2.

The N-terminal domains of all PSG, with the exception of PSG 1, 4 and 8, contain an RGD (Arg-Gly-Asp) tripeptide. This motif is the minimum functional unit for the binding of a variety of adhesive proteins of the extracellular matrix and the blood, such as fibronectin, laminin, tenascin and collagen, to their receptors, the integrins. While this motif can arise without a functional role, its presence in the N domain provides a basis for the hypothesis that it may contribute to the biological activity of the protein. RGD-mediated events include interactions with the immune system, cell recognition, cellular architecture type interactions, and tumour metastasis (reviewed in Springer, 1990 and Ruoslahti and Piersbacher, 1987). The absence of the RGD tripeptide in some PSG suggests that they may have different biological roles.

The A and B domains contain cysteine residues at conserved locations that have the potential to form disulphide bonds and make a looped secondary structure, characteristic of members of the immunoglobulin family (Paxton *et al.*, 1987, Zheng *et al.*, 1990). The N domain, in contrast, does not have these conserved cysteine residues but retains the potential to fold, through interactions between hydrophobic residues at the predicted sites, and by a conserved salt bridge (Bates *et al.*, 1992). The N domain is similar to the Ig-variable (IgV) type domains and the A and B domains to the Ig-C2 type constant domains (Williams,1987).

**Table 2. PREDICTED AMINO ACID SEQUENCES FOR THE C-TERMINAL DOMAINS OF REPORTED PSG TRANSCRIPTS.**

| Subgroup 1 genes | Alternative C-terminal domains | | | |
|---|---|---|---|---|
| | **Ca** | **Cb** | **Cc** | **Cd** |
| PSG1 | DWTVP | EAL | AYSSSINYTSGNRN | GKWIPASLAVGF |
| PSG4 | ..IL. | na | na | na |
| PSG7 | ..SL. | .S. | ...G........D. | ............ |
| PSG8 | ...L. | ... | .........AVY | ..R..V.....I |

| Subgroup 2 genes | Alternative C-terminal domains |
|---|---|
| | **Cm/n** |
| PSG2 | ASTRIGLLPLLNPT |
| PSG3 | .PSGT.H..G...L |
| PSG5 | .PSG..R......I |

| Subgroup 3 genes | Alternative C-terminal domains | | |
|---|---|---|---|
| | **Cr** | **Cs** | **Cw** |
| PSG6 | ETASPQVTYAGP NTWFQEILLL | GPCHGNQTESH | na |
| PSG11 | na | .....DL...ES | GKWIPASLAVGFYVESIWLSEK SQENIFIPSLCPMGTSKSQILL LNPPNLSLQTLFSLFFCFLMAD LVSGLKKVGRGLYQP |

dots indicate sequence identity .
na : sequence not available.

The C-terminal domains vary from 3 to 81 amino acids. The majority are around 12 residues long. The exception is PSG11w which has an unusually long Cw domain of 81 residues. While most PSG are secreted, PSG11w is retained within the cell (Chen *et al.*, 1993). The C-terminal domains also vary in their hydrophobicity, e.g., PSG1d, 7d and 8d have hydrophobic Cd domains and PSG1c, 7c and 8c have hydrophilic Cc domains. The 22 amino acid Cr domains of PSG11 and PSG6 are hydrophobic while the 12 and 11 amino acid Cs domains of PSG11 and 6 respectively are hydrophilic.

The significance of the variability at the C terminus of PSG proteins is unknown. The cell adhesion molecule N-CAM has three splice variants with different C termini. One variant is a glycosyl phosphitidyl inositol (GPI)-anchored form (ssd N-CAM) while the other two are transmembrane forms (sd N-CAM and ld N-CAM). The GPI-anchored form is targeted to the apical surface of polarised epithelial cells, whereas the sd and ld forms are expressed on the basolateral surface (Powell *et al.*, 1991). The different isoforms then, determine the cellular destination of a particular N-CAM molecule. While a similar role for the variable C-termini of PSG molecules is hard to imagine, as the majority are secreted, it is possible that particular C-termini are involved in, or modify, particular interactions with particular cells in some way.

***Genomic structure and organisation of the PSG***

There have been many estimates of the number of genes in the PSG subfamily, based on a variety of methods including hybridisation cloning and sequencing of the N domains (Thompson *et al.*, 1989. Thompson *et al.*, 1990) as well as detection by PCR in blood and placental tissue (Khan *et al.*, 1992, Wu *et al.*, 1993). Most recently, a high resolution physical map has been constructed from overlapping cosmid clones, that spans the region of chromosome 19 which contains the CEA/PSG genes (Olsen *et al.*, 1994). This study has identified a total of twenty-nine genes belonging to the CEA/PSG family, including the PSG genes, genes for CEA, non-specific cross-reacting antigen (NCA), biliary glycoprotein-1 (BGP) and a number of CEA gene family members (CGM1-18). A centromeric cluster of six genes (cen- CGM10-CGM7-CGM2-CEA-NCA-CGM1), spanning 200kb, is separated by 560kb from cluster of twenty-three genes (BGP-CGM9-CGM6-CGM8-CGM12-PSG3-PSG8-

CGM13-PSG12-PSG1-PSG6-PSG7-CGM14-PSG13-CGM15-PSG2-CGM16-PSG5-PSG4-CGM17-PSG11-CGM18-CGM11-tel), which spans 860kb.

All PSG genes characterised to date contain at least six exons encoding a 5'UTR and the first 21aa of the leader peptide (L, exon 1), the rest of the leader peptide and the N domain (L/N, exon 2), an A domain (A1, exon 3), a B domain that is never translated (B1, pseudoexon 4), a second A domain (A2, exon 5) and a second B domain (B2, exon 6) (FIG.2). There is an exact correlation between the protein domains and the exons. Following the B2 domain exon are a variable, gene-dependent number of exons encoding alternative C-terminal domains and 3' UTR that are selected by alternative splicing.

It is not clear whether the difference in the number of central domains between type I (L-N-A1-A2-B2-C) and type II (L-N-A1-B2-C or L-N-A2-B2-C) proteins arises by alternative splicing or from genes with defective A1 or A2 exons (Chou and Plouzek, 1992).

It is clear, however, that the C-termini are alternatively spliced. PSG members can be divided into three different subgroups (1, 2 and 3) according to the organisation of their C-terminal domains (FIG.2).

The genomic organisation of the C-terminal domains of subgroup 1 genes has been determined independently for PSG1, 4, 7 and 8 (Lei *et al.*,1992, Thompson *et al.*, 1990, Leslie *et al.*, 1990, Oikawa *et al.*, 1988). The PSG1 locus appears to produce five PSG, each containing a unique C terminus (PSG 1a, 1b, 1c, 1d, 1e, ). These transcripts are generated by alternative splicing and the use of alternative polyadenylation signals. The Cd domain exon lies adjacent to the B2 exon in all PSG genes. However only transcripts from subgroup 1 genes (PSG 1,4,7, and 8) appear to use this exon .

The genomic organisation for PSG5, a subgroup 2 gene, has been determined also (Thompson *et al.*, 1990, Oikawa *et al.*, 1989a) and the Cm/n region lies about 4.3 kb downstream from the end of the B2 domain. It appears the region lies on two exons with the Cm/n domain and the first 44bp of the 3'UTR on the first exon and the remainder more than 2kb further downstream (Thompson *et al.*, 1990). Only the first exon containing the Cm/n domain has been sequenced and very little of the intervening intron sequence is available. The second exon is predicted from cDNA sequence. The Cm and Cn regions differ only in a deletion in the 3'UTR.

The PSG 6, 11 and 12 are subgroup 3 genes. The cDNA PSG6s, PSG6r (Zimmerman *et al.*, 1989, Barnett *et al.*, 1990, Zheng *et al.*, 1990), PSG11s and PSG11w (Arakawa *et al.*, 1991, Brophy *et al.*, 1992, Zheng *et al.*, 1990, Chan *et al.*, 1990) have been isolated and characterised.

The PSG 12 gene is unreported although a pseudogene, PSG12ψ, has been characterised (Lei *et al.*, 1993). A region that could encode a Cr domain was identified approximately 3.3kb downstream from the end of the B2 domain. A region that could encode a Cs domain was not located but was predicted to lie more than 5kb downstream from the end of the B2 domain.

### *Post translational modifications*

The PSG are highly glycoslyated proteins with carbohydrate contents ranging from 28% to 32%. (Watanabe and Chou, 1988b). It is thought that the N-domain glycan moieties may be essential for the stability of the PSG protein (Chou and Plouzek, 1992).

Sequence analysis of the deduced PSG protein also indicates the PSG may be highly phosphorylated proteins . The proteins contain consensus sequences for phosphorylation by casein kinase II and protein kinase C. There is also a consensus sequence present in the B2 domains of all PSG, for tyrosine kinase. No experimental work has investigated the phosphorylation of PSG proteins.

### *PSG synthesis and expression*

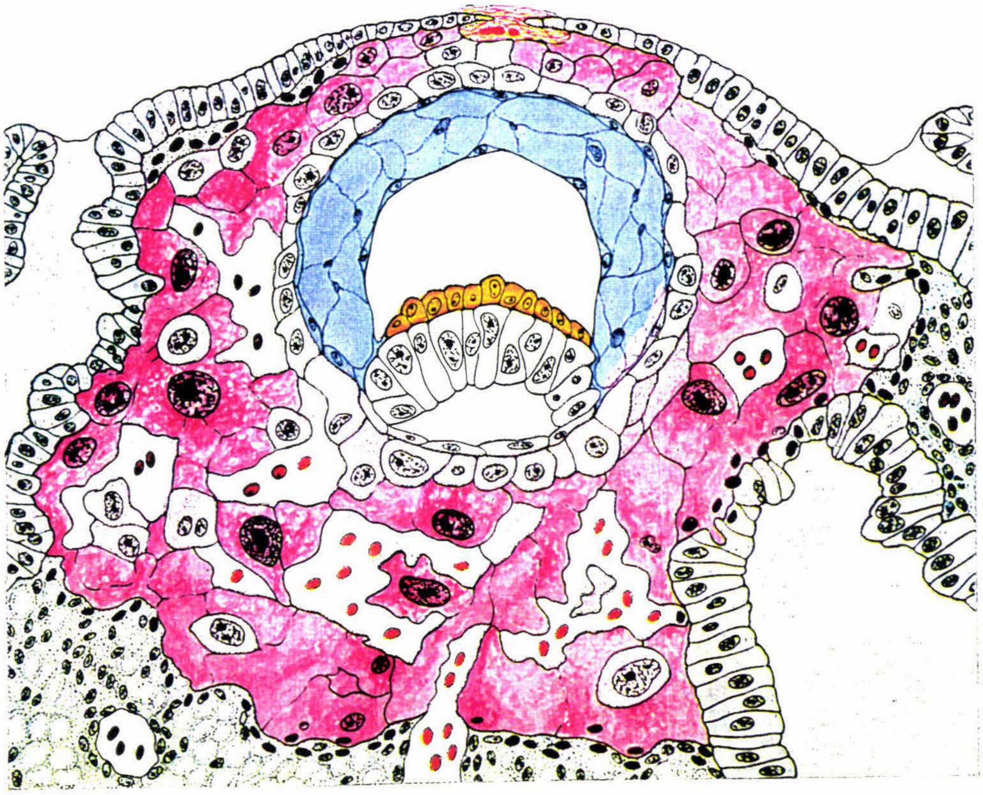
The primary site of synthesis of the PSG is in the syncytiotrophoblast cells of the placenta. These cells originate from the cytotrophoblast cells and form a ring around the developing blastocyst subsequent to implantation. By twelve days, the syncitium is full of large vacuoles or lacunae which form an interconnecting network. Syncitial cells penetrate into the endometrium and erode the endothelial lining of the surrounding maternal capillaries. As a consequence of this, the lacunae fill up with maternal blood, setting up the uteroplacental circulation system by day thirteen (see FIG.3). By the end of the third week primary villi have developed. Lacunae are lined with syncitiotrophoblasts; these are the cells that have direct contact with the maternal blood and are intimately involved with the foetal/maternal exchange.(Ramsey and Donner, 1980)

Syncitiotrophoblasts are known to synthesise and secrete a wide variety of products which include oestrogenic hormones (mainly estriol), progesterone, human chorionic gonadotrophin and human chorionic somatomammotropin (Hamilton *et al.*, 1972). The PSG can be detected three to four days post fertilisation and are regulated independently of other products such as hCG.

The PSG are secreted into maternal serum throughout the duration of pregnancy in a pulsatile fashion with the pulse frequency increasing as the pregnancy progresses. Throughout gestation, PSG serum levels rise exponentially to reach a final concentration of 200-400ug/ml, primarily as a consequence of the increasing mass of the placenta. The decay of radioimmuno assayable PSG in post partum serum appears to be faster during the first 24 hours than later, with a first half-life (0-24 hours after delivery) of 20 hours and a subsequent one (after 24 hours) of 72 hours (Huttenmoser et al., 1987).

Three major PSG mRNA can been detected in placental tissue and in  primary cultures of trophoblasts (Chou and Plouzek, 1992). They are 2.3, 2.2 and 1.7 kb in size and each represents a group of transcripts of similar size. Streydio and Vassart (1990) used specific oligonucleotides to detect PSG1a, 1c, 1e, 2, 3 and 11 by hybridisation in human placenta at different stages of gestation. They concluded that throughout the duration of pregnancy, PSG are expressed in a constant way, with no evidence for developmental regulation, at least for the PSG investigated. Some PSG are preferentially expressed however, PSG1a, 2 and 3 were present at higher levels than PSG1c, 1e and 11.

Further evidence for differential expression patterns of individual PSG genes in the placenta has arisen from a recent analysis of PSG gene promoters. Chamberlin *et al.* (1994) characterised the promoters of six PSG genes and have shown that they fall into two classes



Colour scheme:

| | | | |
|---|---|---|---|
| Maternal tissue | Grey | Endoderm | Orange |
| Syncytiotrophoblast | Red | Extra-embryonic mesoderm | Purple |
| Cytotrophoblast | Pink | Maternal blood corpuscles | Red within dark circles |
| Ectoderm | White with black nuclei | | |

**FIG. 3. A schematic representation of a human implantation site at an estimated age of 11-12 days (from Hamilton *et al.*, 1972).**

which differ in their requirement for activator elements. Modulation of PSG gene expression is, at least in part, achieved by the abundance or availability of certain transcription factors and their interactions with both positive and negative DNA elements in the PSG promoters.

There is some experimental evidence for tissue specific expression of PSG. Work done by Leslie *et al.* (1990) indicates that the PSG6r transcript may be expressed only in hydatidiform mole. It may be possible to exploit this and develop mole-specific probes to enable early detection and diagnosis.

While the placenta is the major site of PSG synthesis and expression, PSG proteins are neither female or pregnancy specific. PSG cDNA clones have been isolated from libraries of testis (Borjigin *et al.*, 1990), foetal liver (Khan and Hammarstrom, 1989, Zimmerman *et al.*, 1989), salivary gland and intestine (Zoubir *et al.*, 1990), HeLa cells (Chan *et al.*, 1988a and b), myeloid cell lines (Barnett *et al.*, 1990, Oikawa *et al.*, 1989a), leukocytes, neutrophils and polymorphonuclear cells (Wu *et al.*, 1993). PSG protein synthesis however, has only been demonstrated in primary cultures of placental trophoblasts, fibroblasts, amnion cells and some human tumour cell lines (Chou and Plouzek, 1992).

All PSG except PSG11w are secreted. PSG11w is predicted to have a hydrophobic C terminus of 81 amino acids (Chan *et al.*, 1991). Preliminary studies on the *in vitro* expression of this transcript in eukaryotic cells suggest that PSG11w remains within the cells, and in fact, is not transported to the golgi bodies but remains bound to the endoplasmic reticulum until it is degraded (Chen *et al.*, 1993).

***Clinical applications.***

There are a number of clinical applications for the levels of PSG in pregnant serum but these are not used diagnostically in New Zealand.

PSG levels have been used to diagnose pregnancy but are generally considered to be less sensitive than the current hCG tests. Low levels of PSG can predict threatened abortion (Hertz and Schultz-Larsen, 1986, Masson *et al.*, 1983) or ectopic pregnancy (Sterzik *et al.*, 1989) and other complications such as foetal growth retardation and intrauterine foetal death when taken in conjunction with ultrasound scans (Chou and Plouzek, 1992, Bischof, 1984, Tamsen *et al.*, 1983). High PSG levels in the amniotic fluid correlate with Meckels syndrome (Heikinheimo *et al.*, 1982). Downs syndrome can be predicted with around 78% accuracy when PSG levels are considered in conjunction with hCG, α-fetoprotein and maternal age (Petrocik *et al.*, 1990). PSG levels and hCG concentrations are useful for detecting gonadotrophin induced pregnancy, as arises in artificial fertility programmes (Rosen, 1986).

PSG levels are used as a prognosis index in breast cancer patients (Horne *et al.*, 1976, Fagnart *et al.*, 1985, Wright *et al.*, 1987) and as an indicator to monitor treatment of choriocarcinomas, hydatidiform mole and gestational trophoblast disease (Tatarinov, 1978). The uses of PSG6r as a probe for the early detection of hydatidiform mole and possibly choriocarcinomas is currently in progress (Leslie *et al.*, 1990).

### *The PSG in other animals.*

Proteins similar to the PSG found in humans have been detected in other mammals such as rodents, sheep and cows. They are however, structurally distinct from human PSG (Thompson *et al.*, 1991, Chan *et al.*, 1988c, Turbide *et al.*, 1991).

## 1.3 The CEA Subfamily.

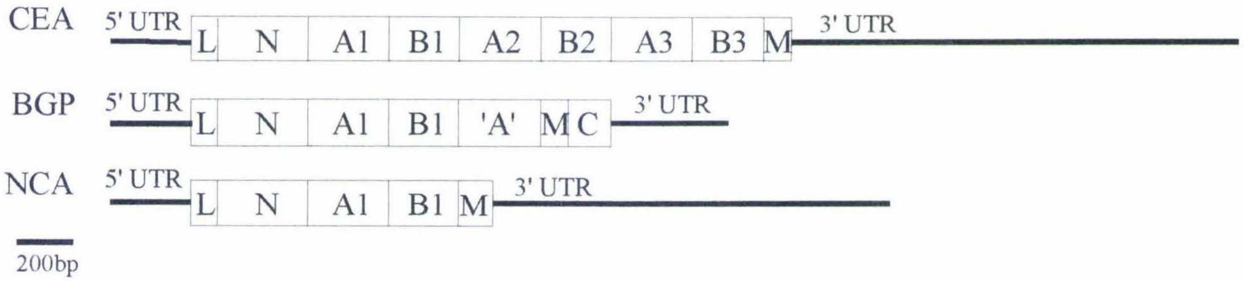
In 1965, Gold and Freedman identified carcinoembryonic antigen as a foetal and colonic cancer antigen present in colonic tumours and foetal gut (Gold and Freedman, 1965). Although CEA has since been detected at low concentrations in a range of normal tissues (Chu *et al.*, 1972 ), in some benign tumours (Kuroki *et al.*, 1984) and in other tumours e.g. breast tumours, it is still one of the most widely used human tumour markers for assessing the treatment of colorectal, breast and lung cancers.

Since the discovery of CEA, other closely related genes and proteins have been identified. As well as the PSG subfamily of proteins, two other CEA subgroup members are well characterised. These are non-specific cross-reacting antigen (NCA) and biliary glycoprotein-1

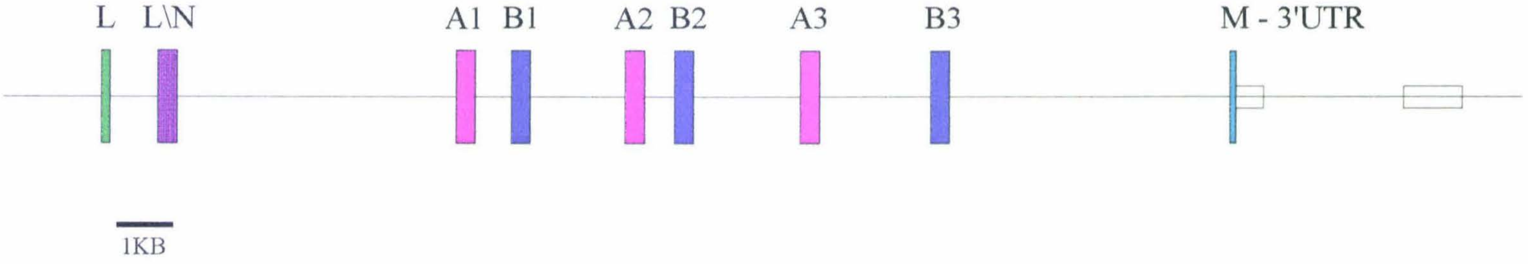
**FIG. 4. Structure and organisation of the CEA subfamily cDNA and the CEA gene.**

The domain structure of three CEA subfamily cDNA is based on deduced amino acid sequence (CEA from Oikawa et al., 1987, NCA from Neumaier et al., 1988 and BGP from Hinoda et al., 1988). The CEA gene (Schewe et al., 1990) contains exons that encode part of the leader peptide (L), the rest of the leader peptide and a complete N-terminal domain (N), central repeat domains (A1, B1, A2, B2, A3, B3) and a transmembrane C-terminal domain (M) and 3'UTR. The 3'UTR sequences are shown as clear boxes on the map. The genes for NCA and BGP are not fully characterised but it is known that the different numbers of central repeat domains occurs as a result of gene organisation rather than alternative splicing. The domain 'A' is a BGP-specific A-like domain and the domain 'C' is a cytoplasmic domain.

## SOME cDNA OF THE CEA SUBFAMILY



## GENOMIC ORGANISATION OF CEA



(BGP) which can be isolated from the colonic mucosa, serum, saliva, bile and faeces of normal individuals (Neumaier *et al.*, 1988, Barnett *et al.*, 1989, Hinoda *et al.*, 1988). Both these protein classes are immunologically crossreactive and share substantial sequence identity with CEA. Unlike the PSG , the CEA, NCA and BGP proteins can be membrane associated.

Other members of the CEA subgroup include a 128-kDa colon tumour associated antigen (TEX) and two meconium antigens of molecular weights 160-kDa and 100-kDa (Neumaier *et al.*, 1988). Whether these are products of uncharacterised genes or isoforms of known genes remains to be determined.

Recently other CEA gene-family member (CGM1-18) genes have been identified (Olsen et al., 1994) but as yet their transcriptional status is not known.

***Primary structure and genomic organisation of the CEA subfamily***

The primary structures of CEA, BGP and NCA proteins as predicted from nucleotide sequence data, are shown in FIG. 4. A short leader peptide of 30-34aa (L) is followed by an N-terminal domain of 108aa. The number of central domains (A-type, 93aa and B-type, 85aa) varies between subfamily members but in contrast to the PSG subfamily, this is as a result of gene organisation and not alternative splicing. As with the PSG, the N domain is Ig-V-like and the central domains are IgC2-like (Williams, 1987). The N domain and central A and B domains of PSG and CEA share around 60% homology (Streydio *et al.*, 1988).

In contrast to the PSG, members of the CEA subfamily are typically membrane bound and have a hydrophobic membrane spanning domain (M) of 26aa at the C-terminus. BGP proteins have a cytoplasmic domain (C) of 71aa in addition to the M domain.

The genomic organisation of CEA is shown in FIG.4. The genomic organisation of NCA and BGP has not yet been fully established.

The structure of CGM13-CGM18 has been characterised by the partial sequencing of these new genes (Teglund and Hammarstrom, 1994). They share identical gene organisation. A single A domain is separated from a single B domain by 0.4kb. These domains share more than 95% identity at the nucleotide level. Approximately 6kb downstream from the end of the B domain is a region that shares 95% identity to PSG C termini and includes exons for the PSG C-terminal domains Ca, Cb and Cc. These CGM genes do not have exons encoding CEA/PSG -like N domains. It is not yet known whether these genes produce functional transcripts.

***Biological roles of CEA subfamily***

CEA, NCA and BGP proteins are highly glycosylated. CEA has 28 potential glycosylation sites and the carbohydrate moiety may comprise up to 60% of the total molecular weight (Thompson *et al.*, 1991).

A biological activity has been established in vitro for CEA, NCA and BGP. CEA and NCA have been shown function in vitro as intercellular $Ca^{2+}$-independent adhesion molecules (Benchimol *et al.*, 1989, Oikawa *et al.*, 1989b) which indicates a possible role in intestinal tissue organisation during development.

CEA and NCA both have a specific affinity for binding certain strains of *E.coli*, by way of a bacterial lectin / CEA carbohydrate interaction (Thompson *et al.*, 1991). Thus CEA may have a role in establishing intestinal flora while NCA on the surface of granulocytes, may facilitate phagocytosis.

Another possible role for CEA on the surface of migrating embryonic cells or metastasising tumour cells, may be in mediating interactions with basement membranes. CEA has been shown to facilitate binding of a colonic adenocarcinoma cell line to collagen type 1 *in vitro* (Thompson and Zimmerman, 1988). Also there is some evidence that CEA may direct metastases from colorectal cancers to the liver by binding a specific receptor on Kupffer cells (cells of the liver, responsible for removing CEA from the circulation) then interacting with CEA on tumour cell surfaces in a homotypic fashion, immobilising them and enabling the establishment of secondary tumours (Thomas and Toth, 1990).

BGP, unlike CEA and NCA, is reportedly $Ca^{2+}$- and temperature-dependent in its binding properties. Sequence homology to a rat ecto-ATPase may indicate a possible enzymatic role for these members of the CEA gene family (Turbide *et al.*, 1991, Rojas *et al.*, 1990).

### 1.4 The Biological Role(s) of PSG

Unlike the CEA subgroup, PSG are predominantly secreted proteins. It is likely therefore that they act through cellular receptors to mediate cellular interactions.

Four potential roles for the PSG have been suggested, they are:

- immunosuppression and the prevention of immune rejection of the foetus (Watanabe and Chou, 1988a, Borjigin *et al.*, 1990)
- an involvement in the invasion of the uterus by the trophoblast (Streydio *et al.*, 1988)
- a role in cellular interactions with the extracellular matrix (Rooney *et al.*, 1988)
- and growth factor activity (Wu *et al.*, 1993, Zheng *et al.*, 1990)

Experimental results, using bulk PSG (i.e. unfractionated PSG isolated from maternal serum) show that the PSG may interact directly with T cells, interfering with their normal interactions and thereby bringing about some form of immunosupression. It has been reported that PSG inhibit E-rosette formation, in a concentration dependent manner, which implies the (T-cell) CD2 : LFA-3 (erythrocyte) interaction is disrupted (Kan and Tatarinov, 1990). The PSG have also been reported to affect the proliferative activity of phytohaemagglutinin-stimulated lymphocytes (Bischof, 1984) and inhibit stimulated lymphocytes in a mixed lymphocyte assay (Zheng *et al.*, 1990).

Since bulk PSG was used for the above experiments, specific interactions could not be identified. The purity of bulk PSG is questionable, especially in the light of contamination by highly potent bioactives such as cytokines and components of the extracellular matrix. Furthermore, the above assays each involve different T-cell receptors so it is not clear whether different PSG bind different receptors or whether they act indirectly, by binding initially uninvolved receptors. One hypothesis, currently under investigation by our group, is that PSG11 binds specific receptors that are involved in the switch to and/or the maintenance of antibody-mediated immunity and suppression of cell-mediated immunity, typical of pregnancy.

The presence of the RGD motif in some PSG suggests a role either in the invasion of the maternal tissue by the trophoblast or in the mediation or co-ordination of cell-cell interactions during embryogenesis (Rooney *et al.*, 1988, Streydio *et al.*, 1988). As discussed previously the RGD tripeptide is present in extracellular matrix proteins, has been shown to be a signal that is recognised by specific cellular receptors and has a role in controlling cell adhesion and cell migration on substrates (Streydio *et al.*, 1988, Ruoslahti and Piersbacher, 1987). Further, there is evidence of some amino acid sequence homology, albeit weak, between the PSG domains and N-CAM, fibronectin and vitronectin, three well characterised cell adhesion molecules (Rooney *et al.*, 1988). A role such as the two described above may well be consistent with the expression of PSG in tumours such as breast tumours, hydatidiform mole or choriocarcinomas.

Some viruses contain the RGD motif in their protein coats and are able to infect human cells through integrin receptors on the cell surface (Bergelson and Finberg, 1993). Perhaps one of the roles of the PSG is to block these receptors, in a non-specific manner during pregnancy.

Another role proposed for the PSG is as a growth enhancer for the cells of the haematopoeitic system. Using reverse transcriptase PCR, Wu *et al.* (1993) were able to isolate most known PSG transcripts from bone marrow and peripheral blood. The levels of expression of PSG transcripts from T lymphocytes were comparable to placental levels. The placenta is known to be a rich source of haematopoietic growth factors and PSG transcripts have also been isolated from foetal liver, a primary site of haematopoeisis in the developing foetus. This potential biological activity is currently under investigation.

## 1.5 Evolution of the PSG and CEA Multigene Family

Like most multigene families e.g. proteins of the blood clotting system, globin genes ( Li and Graur, 1991) collagens (Buttice *et al.*, 1990), the PSG /CEA multi-gene family shows an exact correspondence between protein structural domains and exons (see FIGs 2 and 4). The different domains within each PSG locus also show significant similarity to each other. This indicates both that duplication of domains has occurred and that a single ancestral domain was a precursor in the evolution of this gene family (Streydio *et al.*, 1990).

A possible scenario for the evolution of the PSG family is as follows :

- The precursor (or primordial exon) was A or B domain-like and contained two conserved cysteine (Cys) residues with possibly an Ig-like fold. This molecule may have undergone a series of internal duplications to give a number of similar domains

- One of these domains could have then lost its Cys residues through reduced selective pressure, to become the N terminus

- Then followed a series of further duplications of the A-B unit, and/or rearrangement by unequal crossing over to give a variable number of central domains ,and/or duplication of complete gene units to give multiple genes (Streydio *et al.*, 1990)

Relevant to the question of origins of PSG is the process of concerted evolution. This is a term used to describe the observed homogeneity among members of a gene family, when processes are thought to act to prevent individual members of a family from evolving independently of other members. Thus, the family evolves as a unit, in a concerted fashion. Both the mechanism of unequal crossing over and that of gene conversion are thought to contribute to the process of homogenisation in concerted evolution (Li and Graur, 1991).

It would appear that the CEA/PSG genes have evolved by concerted evolution at least in part, by a mechanism of unequal crossing over . This is suggested by the observation that CEA family members (e.g. CEA, NCA, BGP) each have a different number of central domains (see FIG. 4). These may well be functional products of unequal cross over events. In contrast to the

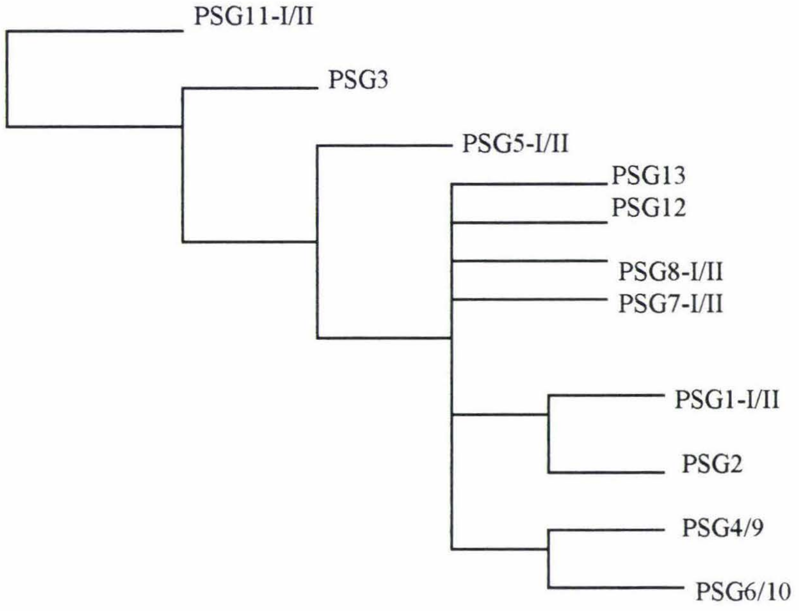


**FIG. 5. Strict consensus of tied optimal trees found under parsimony using N domain sequences (from Khan *et al.*, 1992).**

CEA subfamily, in the PSG subfamily all genes have the same four central domains (see FIG. 2) and different subfamily members appear to have arisen by the duplication of complete loci. This is further suggested from the intron and pseudogene sequences which show extensive similarity also (Rudert *et al.*, 1989).

A multi-gene family encoding proteins similar to the PSG/CEA human proteins has been found in rodents. Early analysis ( Rudert *et al.*, 1989, Thompson *et al.*, 1989) indicated that the two families are non identical and that they have been evolving by parallel gene duplication and subsequent divergence since the mammalian radiation.

The reconstructed evolutionary tree shown in FIG.5 is that found by Khan *et al.* (1992) in a comparative analysis of N-terminal PSG sequences. It was suggested by these authors to represent the evolutionary relationships between PSG genes.

However, the PSG can be divided into three subgroups according to their types of C-termini (see Table 2 and FIG 2). In the tree presented by Khan *et al.* (1992) in FIG. 5, the different subgroups do not cluster together as would be expected if they shared a more recent common ancestor with other members of the same subgroup. In their reconstructed tree some subgroup 1 members join more closely with members of subgroups 2 or 3 (e.g.PSG1 and PSG2, PSG4 and PSG6).

### 1.6 Aims and Objectives.

No complete subgroup 3 PSG gene has been isolated or characterised.

A partial cDNA for PSG3 was isolated by myself and Dr. Mansfield in 1989, from a placental cDNA library (McLenachan and Mansfield, 1989). This cDNA was used by Ms.S.Sims to probe a human genomic cosmid library. Several cosmids were isolated and one of these, hC3.11, was partly characterised by myself and others of our group (Beggs, 1990, Joe, 1994, Joe *et al.*, 1994,) and was found to contain the L, N, A1, B1, A2 and B2 exons of the PSG11 gene. A map of the cosmid hC3.11 showing regions that have been sequenced is presented in FIG.6.

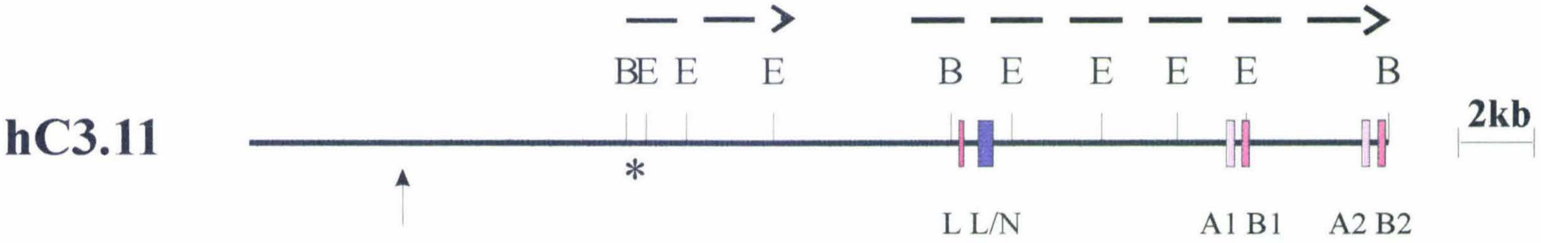


**FIG. 6. A map of the cosmid hC3.11 which contains part of the PSG11 gene.**
Sites for the restriction enzymes B*am*HI and E*co*RI are labelled. The exons for the Leader sequence (L), the N-terminal domain (N), the central domains (A1, A2, B2) and the B1 pseudoexon are shown as coloured boxes. Broken arrows above the map indicate regions of sequence obtained by our group (Beggs, 1990, Joe, 1994 , Joe *et al*, 1994, McLenachan *et al*, 1994). The upstream region of PSG11 that is largely uncharacterised is arrowed. The 0.5kb BamHI - EcoRI fragment used as a probe (see text) is marked ***.**

The primary aim of this project was to complete the characterisation of the PSG 11gene. The isolation, mapping and sequence determination from this locus has now been published (McLenachan *et al.*, 1994). A secondary aim was to examine the evolution of the PSG genes using new sequences from the 3' region of a subgroup 3 gene. The hypothesis of relationship as proposed by Khan *et al.*, (1992) (FIG.5) was tested and this work is currently submitted for publication (McLenachan *et al.*, 1995). An important focus of the evolutionary analysis presented here, has been the attempt to separate historical signals from other (perhaps functional) patterns in the data.

Extensive characterisation of this extremely complex gene family will enable rational investigations into a biological role for PSG to be carried out and may provide further clues as to evolutionary history of the family.