# A novel gene family encoding proteins with highly differing structure because of a rapidly evolving exon

Åke Lundwall*, Claude Lazure**

*Department of Clinical Chemistry, Lund University, Malmö University Hospital, S-214 01 Malmö, Sweden*

Received 14 July 1995

**Abstract Despite vast differences in primary structure, it is here shown that several predominant semen proteins are encoded by genes that belongs to a common family. Members have their transcription unit split into three exons: the first encoding the signal peptide, the second the secreted protein, while the third exon solely consists of 3′ non-translated nucleotides. The first and the third exon are conserved between members, but the second exon is not. The genes for human semenogelins I and II, rat SVSII, SVSIV, SVSV and guinea pig GP1 and GP2 belong to this gene family.**

*Key words:* Semen; Protein; Rapid; Exon; Gene; Evolution

## 1. Introduction

The wide diversity in development of male accessory sex glands results in extensive differences in volume and composition of the mammal seminal fluid.

In rodents, the coagulation gland adds a secretion rich in transglutaminase to semen, catalyzing the slow formation of a rigid copulatory plug by cross-linking of proteins that are abundant in the secretion provided by the seminal vesicles. The predominant clot-forming protein in the rat is the seminal vesicle secretion II protein (SVSII) [1,2]. In the guinea pig, SVP-1 [3,4] serves a similar function, despite an entirely different primary structure [2,5].

In man, the ejaculatory mixing of epididymal sperm and secretions from the accessory sex glands transforms the ejaculate into a loose non-covalently linked gel-like structure. Dissimilar to the covalently cross-linked copulatory plug in rodents, the human gel-like structure dissolves, liquefies and turns into a free-flowing liquid within a few minutes. The major gel-forming proteins in the human ejaculate are semenogelins I and II (SgI and SgII) [6–8]. These are 50–70-kDa proteins present at very high concentrations in secretions from the seminal vesicles. They are encoded by two transcripts that are close to 90% similar in sequence to each other but unrelated to other known proteins.

The rat seminal vesicle secreted protein SVSIV and SVSV are small (~10 kDa) androgen-regulated proteins [9,10]. They are expressed from genes that are organized in a way similar to the semenogelin (Sg) genes, with one exon encoding the signal peptide, a second encoding the secreted protein and a third encompassing 3′ non-translated nucleotides [11,12] It has previously been shown that the 3′ non-translated nucleotides of cDNAs for human SgI and rat SVSIV are similar in sequence [7]. We now extend this finding by showing that these and several other predominant seminal vesicle secreted proteins are encoded by genes belonging to a common family.

## 2. Materials and methods

### 2.1. Southern blots and hybridization

Preparations of DNA from man, African green monkey, cat, sperm whale, cattle, pig, sheep, goat, rat and mouse were digested with the restriction enzyme *Hin*dIII. 3.5 μg of each digest was loaded into a 2.5-mm well on a 10-cm-long 0.7% agarose gel. Electrophoresis was performed at 10 mA until the Bromophenol blue marker had migrated to the anodal end of the gel. The DNA was thereafter partly depurinated, transferred to nylon membrane (Hybond N, Amersham) and fixated by UV irradiation as recommended by the manufacturer. The membranes were prehybridized for 4 h at 60°C in 6 × SSPE, 10 × Denhardt's solution, 0.5% SDS and 100 μg/ml of sheared and denatured salmon sperm DNA. Hybridization was carried out in the same solution for 18 h after addition of probe to yield $2 \times 10^6$ dpm/ml. Following hybridization, the membranes were washed in 2 × SSPE, 0.1% SDS at room temperature and thereafter at 60 °C for 1 h. The following probes were used: a *Bam*HI-*Pst*I fragment from the second exon of the human SgII gene (nucleotides 830–1857) [13]; a cDNA encoding the β-chain of C4b-binding protein [14]; a PCR fragment encompassing the second exon of the rat SVSII gene (nucleotides 313–1546) [2]. The probe was generated from rat DNA by PCR amplification using mixed primers based on sequences flanking the second exon of the SgI, the SgII and the SVSII genes. In a volume of 100 μl, 0.15 nmol of the primers C(A/C)TT(T/C)(C/T)T(A/C)(T/C)T(A/C)TCAATTACCAG and (A/C)(C/A/T)T (T/G)AC(C/A)TTG(C/A)T(A/C)TTGGTC were mixed with 100 ng rat DNA in 20 mM Tris-HCl pH 8.3, 50 mM KCl, 3 mM $MgCl_2$, 0.3 mM dNTPs, 0.1% gelatin and 0.05 U/μl of Taq polymerase. Thirty-five cycles were run for 1 min at 96°C, 1.5 min at 60°C and 2 min plus a 5-s increase per cycle at 72°C. The fragment was purified by agarose electrophoresis, 5′ phosphorylated and cloned into the *Sma*I site of pUC18. Identity of the fragment was confirmed by sequencing of the fragment's ends. Probes were labeled by random priming (Megaprime, Amersham) to a specific activity exceeding $10^9$ dpm/μg.

### 2.2 Computer analysis

Nucleotide sequences were analysed by a set of computer programs from Genetics Computer Group [15]. Sequence comparisons were done by the program COMPARE which generated files that were subsequently displayed by the program DOTPLOT. The comparisons were done with a sliding window of 21 and a stringency set to 14. Multiple sequence alignments were done by the program PILEUP, with a penalty of 4 for the introduction of gaps and a gap length penalty of 0.2.

## 3. Results

It has previously been reported that the 3′ non-translated nucleotides of cDNAs for human SgI and SgII and rat SVSIV are similar in sequence [7,8]. We can now show that the recently published gene sequences of SgI and SgII [13], display similarities to the nucleotide sequence of several other seminal vesicle

*Corresponding author. Fax: (46) (40) 337043.
E-mail: kemall@maja.mas.malmo.se

**Present address:* IRCM, 110 avenue des Pins ouest, Montreal H2W 1R7, Canada.


*SSDI* 0014-5793(95)01076-9

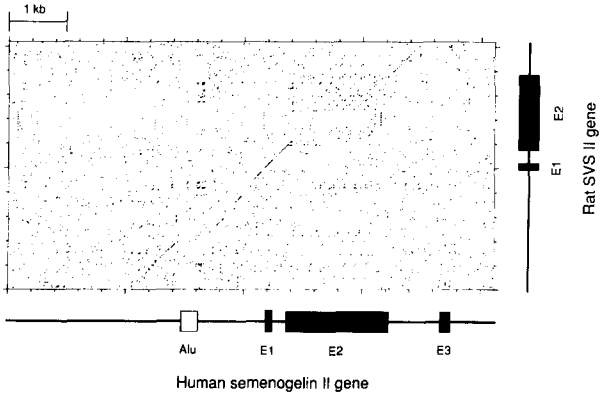


Fig. 1. DOTPLOT showing sequence similarity of the human SgII gene and the rat SVSII gene. The location of exons and an Alu repeat are displayed on the axis as indicated.

transcribed genes as well. Most interestingly, this included the gene for SVSII, the major clot protein of rat semen. Given that SVSII and Sgs are synthesized by the seminal vesicles and constitutes major structural components of semen, they might be considered to serve similar functions, despite the vast difference in primary structure. Fig. 1 shows that the SVSII gene and the SgII gene indeed are homologous as displayed by the long stretches of sequence similarity. However, the sequence similarity does not include the second exon that encodes the secreted protein in its entirety. Therefore, although the human SgII gene and the rat SVSII gene are homologous, they do not give rise to proteins that are similar in primary structure.

Fig. 2 shows sequence conservation of the first and the third exon of several seminal vesicle secreted proteins. Apart from the Sg genes and the SVSII gene, this includes sequences of the genes for the small androgen-regulated proteins SVSIV and SVSV from the seminal vesicle of the rat. As for the Sg and the SVSII genes, the structure of the second exon differ despite the conserved sequence of the first and third exon. Thus, there seems to be a whole family of seminal vesicle transcribed genes, with a common origin, but that give rise to proteins with highly differing primary structure. Members of this gene family have their transcription unit split into three exons, with conserved nucleotide sequences in the first and the third exon, thereby preserving structures of importance for signaling, such as upstream promoter elements, signal peptide and 3′ non-translated nucleotides. In contrast, the second exon with most of the coding nucleotides has undergone such a rapid evolution that there are no sequence similarity.

The major clotting protein, SVP-1, of guinea pig semen is derived from a poly-protein translated from a predominant seminal vesicle transcript. Cloning and sequencing of cDNA for the poly-protein, GP1, and another major transcript, GP2, has shown that they differ in primary structure from SVSII despite their similar function and site of synthesis [5,2]. However, the sequence of the GP1 and the SgI signal peptide have been reported to share 11 out of 15 amino acid residues [5]. In Fig. 2, it is demonstrated that these residues are encoded by conserved nucleotides encompassed by the first exon of the Sg and SVS genes. In contrast, no such sequence conservation is evident in the exon 2 region. Thus, it is likely that the GP1 and GP2 genes of the guinea pig belongs to the same gene family as the Sg and SVS genes.

If the second exon has undergone such a rapid evolution as suggested by the sequence analysis, then a DNA probe encoding the second exon would fail to detect homologous genes in animal species other than those that are very closely related. Hybridization experiments were, therefore, undertaken using probes derived from the second exon of the SgII and the SVSII genes. Fig. 3 shows the result of a hybridization experiment performed at low stringency to a panel of mammalian DNA. The SgII probe recognize restriction fragments in DNA from man and monkey, while other mammals do not hybridize under the conditions used. Furthermore, the number of hybridizing fragments in monkey DNA as shown in Fig. 3A as well as in experiments using other restriction endonucleases indicates that the genome of African green monkey, like the human genome, contains two Sg genes. Therefore, the whole Sg gene locus seems to be highly conserved between these species, contrasting its absence in non-primate mammals. The second hybridization experiment, undertaken with a probe for the rat SVSII gene shows hybridization to restriction fragments in DNA from rat and mouse only, indicating that the translation product might be unique to the lineage Muridae. To serve as a control of the hybridization experiments a third filter was probed by a cDNA for the $\beta$-chain of the human C4b-binding protein. The conservation in nucleotide sequence of cDNAs for bovine and rat relative to human $\beta$-chain is 79 and 76% (A. Thern, pers. commun.). As expected, this probe hybridize to restriction fragments in DNA from all animals in the panel (Fig. 3C).

## 4. Discussion

By sequence analysis and hybridization experiments, we have shown that the second exon of the human SgI and SgII genes and the rat SVSII, SVSIV and SVSV genes has undergone a very rapid evolution. It is likely that genes homologous to the Sg and the SVS genes are present in most mammals and that these genes constitute a new gene family that yields proteins of highly differing structure because of the rapid evolution of a major coding exon. The sequence conservation and site of synthesis, suggests that the genes encoding the guinea pig GP1 and GP2 belongs to this gene family as well.

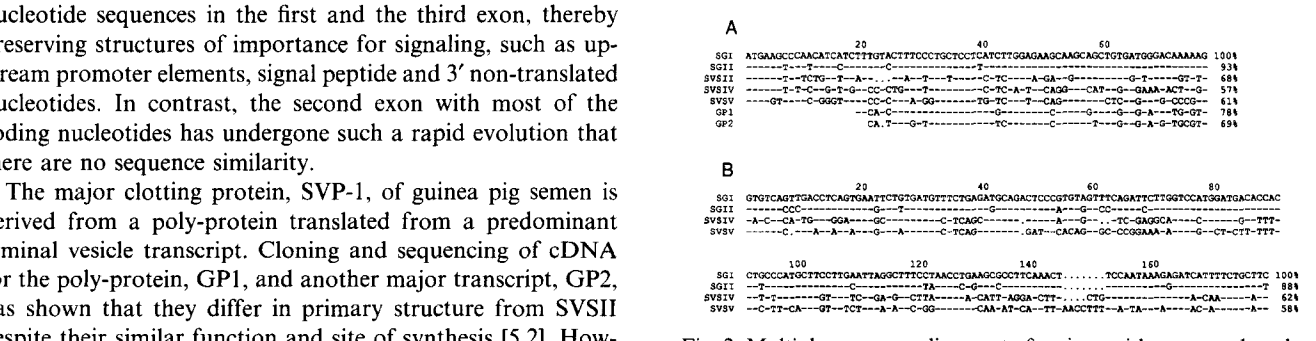


Fig. 2. Multiple sequence alignment of regions with conserved nucleotide sequences. The sequence of SgI is written in full and nucleotides conserved in the other sequences are denoted by dashes. The percent of conserved nucleotides, relative to the SgI gene, is given at the end of the sequences. (A) Alignment of coding nucleotides in the first exon of human SgI and SgII, rat SVSII, SVSIV and SVSV and 5′ ends of cDNAs encoding guinea pig GP1 and GP2. (B) Alignment of nucleotides in exon 3 of SgI, SgII, SVSIV and SVSV.

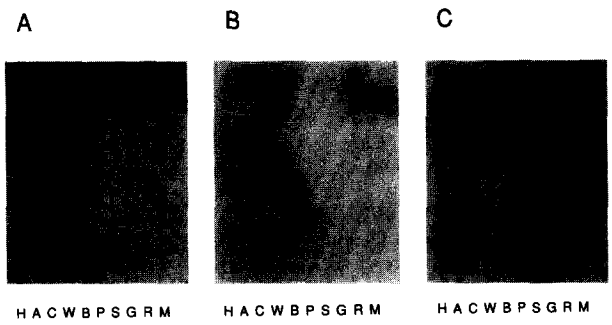


Fig. 3. Hybridization to DNA from different mammals at low stringency. Southern blots of 3.5 μg *Hin*dIII digested DNA was probed by: (A) a *Bam*HI-*Pst*I fragment from the second exon of the human SgII gene; (B) a PCR fragment encompassing the second exon of the rat SVSII gene; (C) a cDNA encoding the β-chain of C4-binding protein. The letters denotes DNA from man (H), African green monkey (A), cat (C), sperm whale (W), cattle (B), pig (P), sheep (S), goat (G), rat (R) and mouse (M).

What is then the mechanism behind the rapid evolution of these genes? Simple point mutations is a main reason behind the the accumulation of sequence variation in proteins from different animal species. However, because of structural constrains, the coding sequences do not accept mutations as readily as non-coding sequence, leading to greater sequence differences in intron than in exon sequences. In the Sg and the SVS genes, the first and, to some extent, the non-coding third exon obey this rule, while the second exon display a reverse pattern with greater differences than in the intron sequences.

More than 80% of SgIs and SgIIs primary structure consists of 60 amino acid residues repeats located in tandem [7,8]. The repeats have been divided into three groups based on sequence similarities, but there are some structural conservation between the groups as well, suggesting that they have a common origin. Probably, they have evolved from shorter repeats located in the N-terminal of the protein. At the DNA level, this appears as 10–20 bp of conserved nucleotides in the 5′ part of the second exon. The highly repetitive structure of the Sgs suggests that the formation of these proteins involved a mechanism of repeated duplications.

The major clotting protein in rat semen, SVSII is constructed from tandem repeats as well [2]. More than half of the molecule consists of repeats that are 13 amino acid residues long. The repetitive nature of this protein suggest that it has evolved through a mechanism of repeated duplications like the Sgs. As pointed out above, there are some 80 nucleotides at the 5′ end of the second exon of the SVSII gene that are conserved in the Sg genes. We suggests that this represents a progenitor of the second exon, that after the separation of the rodent and primate lineage became extended to the size found in the present day Sg and SVSII genes.

Even though the gene structure of the major clotting protein in guinea pig semen is not known, it is likely that a similar scheme of evolution could be applied as for the SVSII and the Sg genes. The transcript GP1 encodes several repeats of 24 amino acids. The clotting proteins from rat and guinea pig are both substrates for transglutaminase but have very differing primary structure. Despite this, our results suggests that they probably are encoded by homologous genes, lending support to the view that the guinea pig is not to be considered as a rodent.

The rat SVSIV and SVSV proteins differ from the clotting and gel-forming semen proteins by virtue of their small size. At the gene level, this is seen as small size and unique nucleotide sequence for the second exon, except for some 25 nucleotides at the 5′ end that are conserved between the SVSIV and V genes. The sequence of these nucleotides are not related to the conserved 5′ end of the second exon of the Sg and SVSII genes. However, the genes for SVSIV and SVSV have a first intron that is double in size to those of the Sg and SVSII genes and when aligned, the intron sequences of SVSIV will continue to be similar to the Sg and the SVSII genes even in the exon sequence for some 80 nucleotides, i.e. up to the point where the sequence of the Sg and the SVSII genes start to diverge (Fig. 4). Therefore, it might be postulated that difference between the SVSIV and the Sg/SVSII transcripts is caused by different selection of splice site to create different second exons from a common ancestral gene. Unfortunately, there are no sequence data available for this region of the first intron of the SVSV gene, but future data from this gene as well as from other genes of the same gene family will either confirm or reject the hypothesis.

Our results suggests that the rapid evolution of gel-forming and related seminal vesicle transcribed genes have evolved by two different mechanisms to yield proteins with highly different structure. The genes have conserved exons for signal peptide and 3′ non-translated nucleotides and the size of the two introns, when added, yields an approximately equal total intron size. The highly varying sequence of the transcripts is caused by the extension of a short progenitor exon by a mechanism probably involving repeated duplications and/or different selection of splice sites. From this we can also devise an evolutionary pathway for the rat SVS genes and the human Sg genes. Because the sequence of the second exon of the SI and SgII genes are as similar to each other as are the rest of the Sg genes they must have been created by a relatively late duplication, probably after the split of the primate and rodent lineages. As pointed out earlier, because of the structural conservation of the first intron and the 5′ end of the second exon, it is likely that Sg genes and the SVSII gene evolved from a common ancestor. In contrast, the SVSIV and SVSV genes carries a first intron that is longer and they also share some 25 nucleotides at the 5′ end

```
  SgI  ccttcttattatcaattaccagGTGGATCAAAAGGCCGATTACCAAGTGA  369
 SgII  -a--tc----------------ag----------------A---G-----C-G  368
SVSII  -----c-cc-c---------ag----GA-------T-AC--C-AG--CAG  362
SVSIV  ga--------g--c-g--ta--cg-ag-agc----a-ag-cg--ggacat  309

  SgI  ATTTTCCCAATTTCCACACGGACAAAAGGGCCAGCACTATTCTGGACAAA  419
 SgII  --C---------------T-----------------------T--------  418
SVSII  C-CA--AGGG---ATG-TT--T--G--A----AC-T-A---T--G-TC-  412
SVSIV  ..caag--t-----tgacta----g---at----g-g-.----a----gc  356

  SgI  AAGGCAAGCAACAAACTGAATCCAAAGGCAGTTTTTCTATTCAATACACA  469
 SgII  ---A-C-A-----T---A-----------------------C-----  468
SVSII  ----AGGAAGTG-GGAA-C-G-TG---AA--CA---TC--G----CACA-  462
SVSIV  c-aaa-g----attg-acc-ca-g-ta-gttac-aatc-cag-gaagc--  406

  SgI  TATCATGTAGATGCCAATGATCATGACCAGTCCCGAAAAAGTCAGCAATA  519
 SgII  ----------CAT----------TG-A-------------------  518
SVSII  C-C--GA-GTTC-G-C-G---GG--GTG-CATGGCGC-G-CAAGTGTT-C  512
SVSIV  ctggt--cttgcagtgtcctgat-t--a-taa--tcgggctat--ag-a-  456

  SgI  TGATTTGAATGCCCTACATAAGACGACAAAATCACAACGACATCTAGGTG  569
 SgII  ----------------------G----------A---A---C-------  568
SVSII  ACAAGA-C--A-AGGTGTA---GG-G-CGCGATTTGT--TA-AGG-CAA-  562
SVSIV  catcca-gta-ggaacagatcatgcttttctctcgtcag-A-AA--CTCA  506
```

Fig. 4. Multiple sequence alignment of the intron/exon boundary at the 5′ end of the second exon of the Sg and the SVSII genes and the first intron of the SVSIV gene. Exon sequences are written in capital letters and introns sequences in lower case letters. A dash indicates the same nucleotide as in the SgI gene. Splice acceptor signals are written in bold.

of the second exon that is not present in the other genes. Therefore, it is very likely that a progenitor of the SVSIV and SVSV genes separated from the progenitor of the Sg genes and the SVSII gene.

An organization of the transcription unit like those encoding the Sg and the SVS proteins, with the signal peptide and the 3′ non-translated nucleotides on separate exons, permits the selection of any nucleotides between these two exons to create a secreted protein without affecting potentially important 5′ and 3′ regulatory nucleotides. A progenitor gene could, therefore, have encoded a protein that is structurally very different from the Sg and the SVS proteins. In these seminal vesicle transcribed genes, the mature protein is encoded by a single exon. However, in principal the mature proteins could equally well have been encoded by several exons and still have the same pattern of evolution. Thus, perhaps does the result presented in this report show a mechanism whereby a rapid process can recruit new coding entities from preexisting genes by selection of new splice sites and amplification of small DNA segments. Thereby, conserving potentially important regulatory nucleotides in the gene's 5′ and 3′ parts while at the same time a new protein is created.

*Acknowledgements:* We thank Dr. A. Hillarp for providing the C4b-binding protein cDNA probe and Dr. A. Thern for providing unpublished results about sequence similarities between human, bovine and rat C4b-binding protein. This work was supported by a grant from the Swedish Medical Research Council (Project 08660).

## References

[1] Wagner, C.L. and Kistler, W.S. (1987) Biol. Reprod. 36, 501–510.
[2] Harris, S.E., Harris, M.A., Johnson, C.M., Bean, M.F., Dodd, J.G., Matusik, R.J., Carr, S.A. and Crabb, J.W. (1990) J. Biol. Chem. 265, 9896–9903.
[3] Notides, A.C. and Williams-Ashman, H.G. (1967) Proc. Natl. Acad. Sci. USA 58, 1991–1995.
[4] Veneziale, C.M. and Deering, N.C. (1976) Andrologia 8, 73–82.
[5] Hagstrom, J.E., Harvey, S., Madden, B., McCormick, D. and Wieben, E.D. (1989) Mol. Endocronol. 3, 1797–1806.
[6] Lilja, H. and Laurell, C.-B. (1984) Scand. J. Clin. Lab. Invest. 44, 447–452.
[7] Lilja, H., Abrahamsson, P.-A. and Lundwall, Å. (1989) J. Biol. Chem. 264, 1894–1900.
[8] Lilja, H. and Lundwall, Å. (1992) Proc. Natl. Acad. Sci. USA 89, 4559–4563
[9] Higgins, S.J., Burchell, J.M. and Mainwaring, W.I.P. (1976) Biochem. J. 158, 271–282.
[10] Ostrowski, M.C., Kistler, M.K. and Kistler, W.S. (1979) J. Biol. Chem. 254, 282–390.
[11] Harris, S.E., Mansson, P.-E., Tully, D.B. and Burkhart, B. (1983) Proc. Natl. Acad. Sci. USA 80, 6460–6464.
[12] Williams, L., McDonald, C. and Higgins, S. (1985) Nucleic Acids Res. 13, 659–672.
[13] Ulvsbäck, M., Lazure, C., Lilja, H., Spurr, N.K., Rao, V.V.N.G., Löffer, C., Hansmann, I. and Lundwall, Å. (1992) J. Biol. Chem 267, 18080–18084.
[14] Hillarp, A. and Dahlbäck, B. (1990) Proc. Natl. Acad. Sci. USA 87, 1183–1187.
[15] Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. 12, 387–395.