Glycoconjugate Journal (1996) **13:** 709–716

Molecular cloning and characterization of the gene encoding rat submandibular gland apomucin, *Mucsmg**

EARL F. ALBONE[1], FRED K. HAGEN[1], CLAUDE SZPIRER[2] and LAWRENCE A. TABAK[1‡]

[1]*Departments of Dental Research and Biochemistry, School of Medicine and Dentistry, University of Rochester, 601 Elmwood Ave., Box 611, Rochester, NY 14642, USA*
[2]*Laboratoire de Génétique, Université Libre de Bruxelles, Brussels, Belgium*

Received 15 August 1995, revised 4 November 1995

Mucin glycoproteins are a major constituent of salivary secretions and play a primary role in the protection of the oral cavity. Rat submandibular glands (RSMG) synthesize and secrete a low molecular weight (114 kDa) mucin glycoprotein. We have isolated, partially sequenced, and characterized the gene which encodes the RSMG apomucin. The gene is encoded by three exons of 106 nt, 69 nt, and 991 nt, separated by introns of 921 nt and 12.5 kb. CAAT and TATA elements are present, at −68 and −26, respectively, in the 5′ flanking sequence of the RSMG apomucin gene. The tandem repeat domain present in exon III consists of ten tandem repeats of 39 nt encoding the consensus sequence PTTDSTTPAPTTK. Sequence comparison and organization of the nucleic acid sequence encoding the tandem repeats of two alleles for this gene suggests that the apomucin gene has undergone recombinational events during its evolution. No significant sequence similarity was found with other mucin genes, or with other known salivary gland-specific genes. The gene was localized to rat chromosome 14 using somatic cell hybrids that segregate rat chromosomes. Since this, to our knowledge, represents the first RSMG mucin gene cloned, we have designated this gene *Mucsmg*.

Keywords: mucins, O-glycosylation, gene-expression

Abbreviations: RSMG, rat submandibular gland; RSM, rat salivary mucin; GRP, glutamine-glutamic-acid rich protein; nt, nucleotide; kb, kilobase

Introduction

Mucin-glycoproteins (mucins) are a principal organic constituent of the mucus secretions which coat the gastrointestinal, respiratory, and urogenital tracts. This slimy, viscoelastic coat aids in the protection of these exposed epithelial surfaces from microbial and physical insult [1]. Previous studies have shown that the mucin polypeptide backbone (apomucin) usually consists of tandem arrays of repeating amino acid sequence rich in threonine, serine, and proline, to which are attached numerous O-linked oligosaccharides. These O-linked side chains may constitute as much as 80% of the molecular mass of the molecule. Salivary apomucins vary greatly in size, with two general classes being identified. 'High' molecular weight salivary mucins are characterized by the presence of a cysteine-rich domain which forms multimeric complexes. In contrast, 'low' molecular weight salivary mucins lack this cysteine-rich domain and remain monomeric.

Rat submandibular glands (RSMG) secrete a low molecular weight (114 kDa) mucin which forms a major component of rat saliva. Conceptual translation of cDNAs encoding this apomucin [2] revealed three distinct regions; a basic N-terminus rich in the amino acids glutamine, proline, and tyrosine, but lacking in hydroxyamino acids, a threonine- and proline-rich tandem repeat segment $(PTTDSTTPAPTTK)_{10-11}$ which showed allelic polymorphism in tandem repeat number, as has been seen for other mucins [3, 4], and a serine-, threonine-rich C-terminus. Although there is no signifi-

*Sequences reported herein have been assigned GenBank accession numbers U33441 and U33442.

‡To whom correspondence should be addressed.

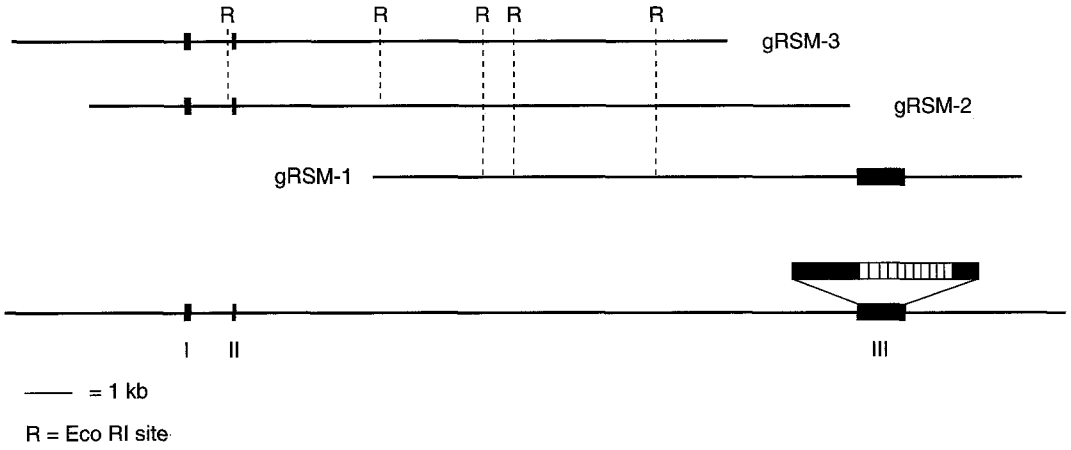
cant sequence similarity between the RSMG apomucin and human salivary mucin MUC7 [5], these mucins share a similar architecture. Therefore, it appears that RSMG mucin represents an analogue to the human salivary mucin, MUC7 [5], which is thought to promote the clearance of bacteria from the oral cavity [1].

Although cDNA cloning has revealed valuable information concerning the primary structure of mucins from a variety of sources, only the genes for the membrane-bound human tumour-associated epithelial mucin *MUC1*, and the mouse homologue, *Muc 1* have been cloned in their entirety [3, 6]. Partial gene structures for the human *MUC2* [7] and the canine tracheo-bronchial mucin (*CTM*) [8] are also available. The promoter for *MUC1* has been characterized, and elements which govern its tissue-specific expression have been identified [9–11]. Previously, by Southern blot analysis of rat genomic DNA, we estimated that the RSMG apomucin was encoded by a single copy gene [2]. In the present study we have isolated and characterized a gene, termed *Mucsmg*, which encodes the RSMG apomucin.

Materials and methods

Isolation of genomic clones of Mucsmg *gene*

To isolate genomic clones corresponding to RSMG apomucin, two screenings of a λDASH II rat genomic library (Stratagene) were performed. The library was plated out at a density of 5–8 × 10^4 phage per 150 mm plate using LE392 as host (Stratagene) (10^6 phage total). The plaques were lifted onto nitrocellulose filters (Schleicher and Schuell), denatured in 0.5 M NaOH, 1.5 M NaCl, neutralized in 1 M Tris (pH 8.0), 1.5 M NaCl, rinsed briefly in 2 × SSC, and baked for 2 h *in vacuo*. Screening was performed using the insert of cDNA clone pRSM-3, available from a previous study [2], which corresponds to position 123–1203 of the full length cDNA clone. Hybridization and washing were performed as previously described [2]. Four rounds of screening were performed, and two clones were isolated. Phage DNA was prepared by the method of Chisholm [12], digested with various restriction enzymes, and subjected to Southern blot analysis [13]. Results indicated that the clone gRSM-1 contained a single exon (exon III), but lacked the remaining 5′ exonic sequence. Therefore, the library was replated at a density of 10^5 phage per 230 mm × 230 mm plate (10^6 phage total). Plaques were lifted onto 23 cm × 23 cm nylon filters (Schleicher and Schuell) and denatured, neutralized, and baked as described above. A 5′ cDNA probe was generated by polymerase chain reaction (PCR) using the following primers: EA-1, TTCTTCTCGAATTTTCAACCGTAGC; PE-2, CGTAAA ATATGAAAGAAGAGCCAACAGG, whose 5′ ends correspond to positions 1 and 174 of the published cDNA sequence [2]. Amplification was done by denaturation for 3 min at 94 °C, followed by 30 cycles of 94 °C, 15 s; 51 °C, 15 s; 72 °C, 30 s; followed lastly by a 72 °C



Splice Acceptor $(Y)_{7\text{-}14}$NYAG/G	Exon consensus	Exon Size (bp)	Exon Position	Splice Donor MAG/GTRAGT	Intron Size (bp)
–	I	106	+1	AAG/GTGAGT	921
CCTTCTTTCAG/G	II	69	+1028	ACG/GTAAGT	≈12,500
TTTTTCTTTTTATTTCCACAG/C	III	991	+14124		

Figure 1. Overlap of genomic clones encoding the RSMG apomucin gene, *Mucsmg*. Exons are indicated by the solid rectangles. The repeat domain in exon III is indicated by an array of open rectangles. *Eco* RI sites are indicated by an 'R'. The sizes and positions of the exons (in nt) and introns are indicated in Table 1. Consensus and actual splice donor and acceptor sequences are also given. Invariant sequences at the splice junction are underlined.

extension for 10 min. The cDNA clone pRSM-2, representing a full length RSMG mucin cDNA clone but containing only two tandem repeats [2], was used as a template in the amplification reaction. Hybridization and washing were done as described above. Two clones were isolated (gRSM-2 and gRSM-3) from three rounds of screening. These clones were digested with *Eco* RI and *Sac* I and were found to overlap with phage gRSM-1. Southern blot analysis using the 168bp PCR-generated screening probe indicated the presence of two additional exons.

DNA sequencing of the Mucsmg *gene*

Restriction fragments of the cloned DNA were subcloned into pBluescript KS(+) and SK(+) and propagated in XL-1 Blue (Stratagene). For sequencing of the fragments containing the promoter and exons I and II, single-stranded DNA from both strands was prepared using helper phage VSCM13 (Stratagene). For sequencing of the fragment containing exon III, double stranded DNA was used as a template. Sequencing was performed by the chain termination method using the TaqTrack Sequencing System (Promega) and [α^{35}S]dATP, as recommended by the manufacturer, using both vector and gene-specific primers. Both strands of the DNA were sequenced. Sequences obtained were analyzed using the software AssemblyLIGN and MacVector (ver. 4.5) (IBI).

Primer extension

Previously, we demonstrated a single start site using an antisense primer corresponding to positions 60–94 of pRSM-2, which corresponds to exon 1 of the *Mucsmg* gene [2]. To rule out the possibility of an alternative start site in intron no. 1, we have used an antisense primer [CGTAAAATATGAAAGAAGAGCCAACAGG] corresponding to cDNA positions (147–174) and genomic map positions 1069–1096, which is located in exon 2. This probe was end-labelled with [γ^{32}P] ATP (6000 Ci nmol^{-1}, Dupont New England Nuclear) and 1.1×10^5 cpm primer used in a primer extension reaction with ~50 μg of RSMG total RNA as template. The resultant product was separated by 6% denatured polyacrylamide gel electrophoresis and signal localized by autoradiography. A standard sequencing reaction was loaded in a parallel line to serve as a size marker.

Chromosome assignment

Cell hybrids used in this study were derived from the fusion of mouse hepatoma cells (BWTG3) with adult rat hepatocytes and have been described previously [14]. These cell hybrids have lost rat chromosomes and have been used to map several rat genes. DNA was extracted and analyzed by the Southern blot method [13] after blotting to Nylon membrane. The probe sequence is shown in Fig. 3.

Results and discussion

Screening of the λDASH II rat genomic library

From a total of 2×10^6 phage using either a nearly complete RSMG mucin cDNA or the 5′-most portion of the cDNA as a probe, a total of three lambda phage were isolated. Southern blot analysis indicated the presence of three exons. Approximately 4 kb of 5′ flanking sequence was isolated.

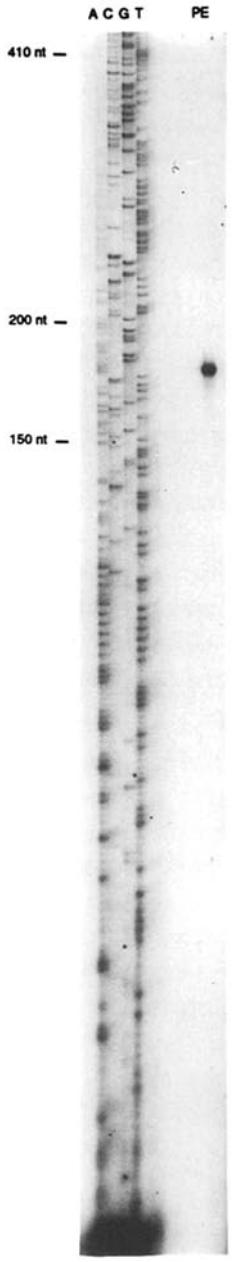


Figure 2. Primer extension analysis from exon II. Lanes A, C, G, and T are a random sequence used as a size marker. PE indicates the lane of the 174 nt PE2 primer extension product. The PE2 priming site is indicated in Fig. 3.

```
                                                GATCTTTTCATTTTTTTTATTTTTTGATAATTA -1914
GATGGCACAACTCTCAAGCACAAATTTTATACAAAACAGTGTGTCAAAAAGAAAAAAGTGTACCTTCCCT -1844
TCAACTTTAGCACTTCCCAAGAAGACCACATACAGCTGGGGCAGAATTATGTACAGCTGAAAGGACCTGG -1774
TTTAACGACCCTTCCCAACACTCATTCTATGATATTATTCTTGTTAGCTGTCAGTAATCACAGCTCTGTT -1704
AAATTAATTGGTAAAATTTTTCTCAGAGAATCAAGTTTCACTTGATGTTGATCCAGATTTATTAGGGAAA -1634
AAAATATCTACTAACTGATCAAATTCACTAATAGCAGTAAGACATTATCAACAACTTGAATTTCCTGTGA -1564
CAAAGACTTAAACATTGTATGAGAATTTATCTAGTCTAGTCTTTCATAGTCAATTTTTTAAGAAATTACT -1494
CTTGACAAGTAATGGCCTTTGTTTTTTAATAAAAGTGTGGTTCTGACTGCTTCCGACATTCTAATCTTTT -1424
GTTTCCTGTGACCCATGAAGCAATTATTTCAATTTGACAGCCTGCATCCTTCTGTCCTCCACAATCAGTG -1354
ATCCATACAAACATAAGTCACTTTAAAACATTTGGCAAAAAATGTATCCAGGCAGCAAAACAATTCTAGA -1284
CCAAGCACAAAGACGGGCTGTTTATCTTTTTAATAAAATGGCCATGTTACTAGAGAACACGCTGTTCAGA -1284
ACATCCCCATGGTTGGCTTAGAAGGTCATGAAAATGGACAACAGGTTGTGCAAAACAATCTGCAAGCCAG -1144
CAACTATGCTGGAGTTCCAGATGCTACATTTCCCACTAGCAACGCTCCAGCAGGTCAGGGGAAATGTGGT -1074
GGCCCAAAGACCAATGTTTCCTATGATTCCATAATTTTAAAGAGGTGGTTACAAGGTCACATGAGCTAAC -1004
CCTTCACGAGTTGTTAGTATACTCTATCACCTGGTGGGTCTAACAGATTTAAGCTATTGGAGAATGAAGT  -934
CTAACATTCAGCAAAAGGGGATTTTCTCCCAAATGAGCAGATTGCATTAGATATGGTTACCTCCCTTTAG  -864
CACCTGGATTTATTTTAAACATAAAATTGTGTAAAATGCTATGATCTCTAAATGTAAATTTAAATTGGAT  -794
TAGGCAAATTTTCTGTAATTACCCTCATTTTTAGGCTAGTGTACATTGTATCCATTCATCTTGTTTTGGA  -724
TAATGAAAATAGCAAGTGTTTTTAGAACCTAAAATATATTTATTATGCTTAAAAGTCCAGCTGTCATGAG  -654
CTACAGAAAAAAATGTGTACAGTAATCCCCATGTTATCCATTTCAGTGACTCAAATGGGTATGCATAGTA  -584
CCTTTAAATAAAATTATGCCTCTAGGAATGAAATATAAAATGAAAGTTAATAATGTGTCTTTCCCTGATT  -514
TTTTTTTCATCTCCCACAACCAGTCTTTCTGCTGTTTCAAAGCATTTTTCCCATTCTCAAAATCCTTCCC  -444
ACAAAATAAATGTCTTCACAAATAAAACAACCCATGTTGTTAATATGTCAGTTCTAATGCATAATTCTAG  -374
CCATGTGGAAGCTCTAAGCTCAACAACATGAATATATCACTAATAAAAATTTAGCTTGTAACACTTGGGC  -304
TCTTTATAAAAGGTATTTTCTCAAATTCTTCCTGGGATAAGGCTTGAGAATGAAATATTATAAATACCCC  -234
ACACACACATAATGAGGATATGCATCTACATATGAACCCTCAACCCTGTCAGCTCCAACCTGTTGGACAG  -164
TACTTGGGTCACAGTGCCGTGGGTGGAGTTGCCAAGGTGAAAGCTGTGACTATACTTCTTTGTTTCAGTA   -94
AACATTATCAATCCATCTTATCTTCTGTCAATCTTACTGTCCCTTTACTAGGAATCAGAGAACAGTATTT   -24
AAGAAAGGATGTGTTCTATGAAAGTTCTCTTTCTTCTCGAATTTTCAACCGTAGCTACCAAAACTGAATA   +47
TTTGGCAAAAGTAAGCTGAAGCAACAGTTGATTGTTCCAAAGGAGAATCTCCCACCAAGGTGAGTACTTG  +117
                                                                PE1
TTCATAAATATATAAGGTTTTATGACTTCACAGTTCTGAGTTTTGCCCTTAGTATGAAGCATATCATTCT  +187
TGTGACTTTTAAGAGAAAGAGCCACAAGGACAAAGCATTTTAGGTGGTAGAGCACACTGAACTTTAATAA  +257
GTAATGAGGGAGAAGGGATATGAAGACCTGGCACACTGCACAATGTAACCCAACTTTTGCAATCTGTTGA  +327
TTTCCCAGCTGATGTATTTTGATGGTCTCTACAGTAAAACATGAGTAGACGTCCTTTGAAGATAGCCATA  +397
AATGTAGATTGGCATTGCTGCCTTTCCTGTTTGAGTTACATTTATTACATATTCTTGAGCTAATTTTAAG  +467
ACAGCACATTAGCAGTGTACAGAGAGGATAACAGCTGCTGCATCTGTAAACAACTCTTCCTATACTTAAA  +537
GAGGGCACAAAGACCCCAAGACCCCTTGTCAACATAATATTCATAGTTGAGCTTTGCATTCCCAGAATCA  +607
TTTCCACAATACTAGGATACGTTGTAGAGTTATTGTCAGTTTAAAAGTCCAAACTTTGACAGCCAATTCA  +677
CATGCTGATTAGCCACAATTGGCTTTTGAGAATCACTTGAAAGGAAGGCTTATGAAAGCAGTAATCATTC  +747
AGTAGACTGATATGAGTAATGAAAGTGGTGATGATAATACCGCAATAATTTTACCAAAATATTATTTTCA  +817
CAATAACCTATCCACACTGGCAATACCTAGGCTGCATGCACACATGGTTATCTTCTCTAAAAATCCTGTG  +887
CCTCACTAAATGTGGTAGCAGCATTTGTTCGTCATGTTACCTGACCACAGTCAGAATTCCTTCCAAGTTT  +957
TCATTTTAAACTAAGGAACATGCTGTTGACTGACTTGACTAACTGTGGTCAATCACTGACCTTCTTTCAG +1027
```

Figure 3. Sequence of the *Mucsmg* gene. Exonic sequence is indicated by single underline. The conceptually translated apomucin is indicated by single-letter amino acids. Numbering of nucleotide positions is relative to the transcriptional start site (position = +1), where positions in sequence containing exon III are based on a size of 12.5 kb for intron II. Positions of potential TATA and CAAT boxes are indicated by a medium thickness underline. The tandem repeats are indicated in brackets [ ]. The putative signal peptide sequence is boxed. Positions of primer extension annealing sites PE1 [2] and PE2 are indicated by horizontal arrows. The exon 3 probe used for chromosomal localization is shown with a thick bold line.

```
GAGCAACATTAAGAAATGAAAAGGGAAACTTTCATCTTGGGCCTGTTGGCTCTTCTTTCATATTTTACGG +1097
                                                                  PE2
                  M  K  R  E  T  F  I  L  G  L  L  A  L  L  S  Y  F  T
TAAGTTCCCCCCAGTAGCCAATATCCTTGATTCACTTATTGTCAAAACTTCAAAAGACTTCTTGTATTTT +1167
TGTTCTTGTTTTTTGTTTGTTTGTCTGTCTTGTTTTTTTTACTAGCAATAGCCTTTATCACTTAAAACAT +1237
GGGTTAGGCTCCTGCTCACTATAGACTTGCTTTGATTTATCTT.............                 +1280

                    intron II approximately 12.5 kb

          .............AAAAAAATAAATTTTAAGATATATGTAATTACTATGTGTTGGTTTTG +13570
AATAAAATAAAATAATCACAAAACATAGTTATATATAATAAAATAATAAGTTTTAATGTTCTACAGCACA +13640
GCAGTGACTATATTTTATAAAAACCATATATGTATTTTATGAGGTGCTCAAAGTCTCCAAACACAAAATT +13810
ATAATAAGAAATATACACATAATTACTGTCATTTGGTCAATGCATGTACTGATATTGCCTATATTTCACA +13880
AATATATACCATATAAATGCAAAAATATATTACTTGAGTTCAGCAAAATATCCTTAAATATTTAATACGA +13950
AATGTTAACTATCCTTAGTTGATCATTATATGTTGTATACTTGTATAGAATTATACTTTATCCCTCTAAC +14020
TATACAATAGAAATAATAATTAATACAAACATAGCCATATGTGCTAAGGTTTGAAATTCAAATGATGGAA +14090
TTTAAAAGTTTCTTTTTCTTTTTATTTCCACAGCCTGGAGAAAGTCATCACTTCCAGCCAAAACCACATC +14160
                                  P  G  E  S  H  H  F  Q  P  K  P  H
CATACCAAAGGCTACAGCAACCCATTTACCACAGACGACATTCACAAGTCTCTTCTATTTACCCAAGATA +14230
P  Y  Q  R  L  Q  Q  P  I  Y  H  R  R  H  S  Q  V  S  S  I  Y  P  R  Y
TGGTCAATATCCACGTTATTTCTATGTTTCACAGAAACAGCAAGCTCAAAAACCTCAAATTTTACCAATT +14300
  G  Q  Y  P  R  Y  F  Y  V  S  Q  K  Q  Q  A  Q  K  P  Q  I  L  P  I
CAAACTCCATGGCAACGTGTCTGCCCTCCAGGATATACTGCGAGGCTGCTCCATTACCATTATTCTAGGT +14370
 Q  T  P  W  Q  R  V  C  P  P  G  Y  T  A  R  L  L  H  Y  H  Y  S  R
TTCTATGCATTCCTAATAAACAGTTAACCTCTGACAAAATAGAAACAAAAGTCACTACACCAGCACAGAC +14440
F  L  C  I  P  N  K  Q  L  T  S  D  K  I  E  T  K  V  T  T  P  A  Q  T
CACCAAGCCTACCACAGATTCAACCACACCAGCACCGACCACCAAGCCTACCACAGATTCAACCACACCA +14510
  T  K][P  T  T  D  S  T  T  P  A  P  T  T  K][P  T  T  D  S  T  T  P
GCACCAACCACCAAGCCTACCACAGATTCAACCACACCAGCACCGACCACCAAGCCTACCACAGATTCAA +14580
 A  P  T  T  K][P  T  T  D  S  T  T  P  A  P  T  T  K][P  T  T  D  S
CCACACCAGCACCGACCACCAAGCCTACCACAGATTCAACCACACCAGCACCGACCAACAAGCCTACCAC +14650
T  T  P  A  P  T  T  K][P  T  T  D  S  T  T  P  A  P  T  N  K][P  T  T
AGATTCAACCACACCAGCACCAACCACCAAGCCTACCGCAGATTCAACCACACCAGCACCGACCACCAAG +14720
  D  S  T  T  P  A  P  T  T  K][P  T  A  D  S  T  T  P  A  P  T  T  K]
CCTACCACAGATTCAACCACACCAGCACCGACCACCAAGCCTACCACAGATTCAACCACACCAGCACCGA +14790
[P  T  T  D  S  T  T  P  A  P  T  T  K][P  T  T  D  S  T  T  P  A  P
CCACCAAGCCTACCACAGATTCAACCACACCAGCACCGACCACAAAAATACCTACTACACCTAAGCCTAG +14860
T  T  K][P  T  T  D  S  T  T  P  A  P  T  T  K] I  P  T  T  P  K  P  S
CACCTCAACAGCCATACCTACATCAACTAACTCTGCTAACAGCTCTTCCTCTACTACTACATCAAGTACC +14930
  T  S  T  A  I  P  T  S  T  N  S  A  N  S  S  S  S  T  T  T  S  S  T
ACCATCCAAACTACAACTCTGTCACCTTTTCAACAGATGCTTCAGTGGCTTCAGATGTACTTTGGTTAAA +15000
 T  I  Q  T  T  T  L  S  P  F  Q  Q  M  L  Q  W  L  Q  M  Y  F  G  *
GTAGGATGGGATGGGATGGTCTCCAAACTGTCAGAACAGTTTTATCTCTTAGGATAAATAAGCCTCAATG +15070
ATGATTTTAAAGAAATCAACCTGATCTTACTAGAAAATCAAACAAATAAAAACAATTTGAGCAATGAAAT +15140
GCATCTCTTTTTGTCCTGATGACTACCATGTTATCCTGTGTTTTACCATCTAACCACCAACTGCTAAATG +15210
GGCTTCCTAGAAGAAATCAGGGAGCCTTTGATTGAAAATACCTGTCTATATACACAGATGTTTACATATA +15280
TATCTATTGTAAGTTGCATGAGTTTAACTAAAAACAAAAAAATACAAGAGCTATTAAGGCATTCAATCTT +15350
GTTCCAATGAAGCTGGTAAGGTACCTAGGCATGGTGGGGCAAAGGGAATATAAAAGTAAATAAATTAAAT +15420
TAAGTAAAATAAGCAGAGAGACCCATCTTCCCTGCTTGCTGAGTCCTCTGGAGAAAGCAGAGCTGCCCTG +15490
AGCAGACTGCTATACCAGGGTGACCTCCATTCTTCCCACACTTCTC                          +15536
```

Structure and sequence of the RSMG apomucin gene, Mucsmg

The structure of the gene encoding RSMG apomucin was determined based upon sequence, Southern blot analysis, and restriction endonuclease digestions (Fig. 1). The gene is comprised of three exons separated by two introns. All intron/exon boundaries were based upon consensus sequences [15] as well as comparison to the cDNA sequence. The first exon comprises the initial 106 nt of the cDNA sequence, which represents nearly the entire 5′ untranslated sequence. Exon II is 69 nt in length and contains the remaining 16 nt of 5′ untranslated sequence, as well as nearly the entire putative signal peptide. Exons I and II are separated by a 921 nt intron, which was sequenced in its entirety. The size of intron II was estimated by restriction endonuclease mapping to span 12.5 kb. Exon III is 991 nt in length and encodes the remainder of the apomucin transcript, and contains the polyadenylation signal AAUAAA 116 nt from the termination codon which corresponds to the site for polyadenylation in the cDNA.

Primer extension

Assignment of the transcriptional start site was initially based upon primer extension analysis performed in a previous study [2], using a primer (PE1) which hybridized to position 61–95 nt of the first exon. This transcription start was confirmed by primer extension using the PE2 primer, which hybridizes to genomic sequence position (+) 1069–1096 nt of exon II (Fig. 2, position in exon is shown in Fig. 3). The 174 nt primer extension product for the PE2 primer (Fig. 2) maps the 5′ end as the message to the identical cDNA position as previously demonstrated for PE1 primer extension experiments, indicating that no additional start site was located in intron I. A TATA-like motif TATTTAA was found at position −26 relative to the start site, which is similar to the consensus sequence TATAWAW, differing only in the fourth position. In addition, a consensus CAAT-like sequence was found at position −68 (Fig. 3) [16].

Comparison of the Mucsmg *sequence with the cDNA sequence*

The sequence of the conceptually spliced exons of *Mucsmg* is identical to the previously published cDNA sequence [2], with the exception of nucleotides at the 5′ end of the cDNA and in the repeat domain. Actual nucleotide sequence in the genomic clone at positions +4 and +6 were found to be changed from T to C, and the +1 position was found to be a G, suggesting that the 5′ end of the cDNA reported earlier [2] either reflects a minor sequence polymorphism or a cDNA cloning artifact during linker ligation.

Two alleles (*A* and *B*) for this gene are represented by a genomic clone of RSMG apomucin, which contains ten tandem repeats (allele *B*), and an eleven tandem repeat motif found in the cDNA clone (allele *A*). The presence of both the ten and eleven tandem repeat alleles were confirmed by the identification of a ten and eleven tandem repeat restriction fragment in a previous Southern blot analysis [2]. The repeat region of the genomic clone contains point mutations identical to those found in the cDNA clone, as well as an additional variant of the repeat sequence, in which an A to G mutation at nucleotide position 7 occurs, resulting in a threonine to alanine change in the primary sequence (Fig. 4). In addition, the positions of the variant repeat sequences with respect to the consensus repeats and to each other differs between the ten and eleven tandem repeat-containing alleles, suggesting that RSMG mucin has undergone numerous recombinational events during its evolution.

These observations are consistent with the current views of mucin gene evolution [17]. It is thought that the tandem repeats are inherently unstable and recombine by unequal crossing over during the evolution of the gene, resulting in the expansion of the repeat motif into enormous domains. A consequence of this recombination is the generation of alleles differing in tandem repeat number. Allelic variation in tandem repeat number has been shown for a number of mucins, including RSMG mucin. The relatively small size of the repeat domain of RSMG mucin, when compared to some of the other cloned secretory mucins, may indicate that the gene is in the early stages of evolution, or, alternatively, that there is a functional selection which does not allow the repeat domain to expand.

Chromosome assignment of Mucsmg

The mucin gene was localized using a panel of 13 rat-mouse somatic cell hybrids that segregate rat chromosomes. A 204 bp fragment corresponding to exon III (Fig. 3) was used as a probe and detected one rat *Hind* III fragment (19 kb), and one mouse fragment (6 kb). The rat fragment co-segregated clearly with rat chromosome 14, as shown in Table 1; at least three discordant clones (gene present/chromosome absent or vice-versa) were counted for each of the other chromosomes. The mucin gene thus resides on rat chromosome 14.

Comparison of the Mucsmg *with other mucin and salivary gland-specific genes*

A search of the Genbank/EMBL rodent/human databases revealed limited sequence similarity between *Mucsmg* and the human proline-rich peptide, *PRP P-B* [18]. The gene products share a 68% amino acid sequence similarity in their signal peptides. No significant similarity was seen

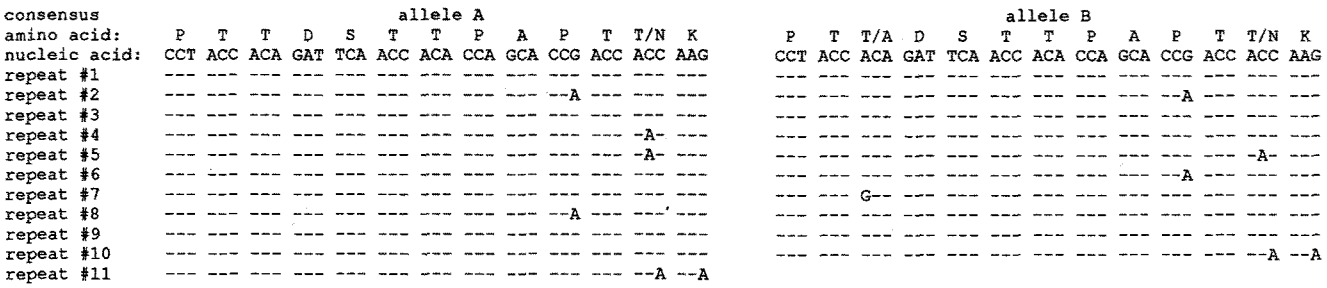


Figure 4. Alignment of repeat sequences of allele *A* and *B* for the RSMG apomucin gene with the consensus repeat sequence. Dashed lines indicate position of a conserved base; only diverging sequences are shown. The nucleotide change in repeats 4 and 5 of allele *A* and repeat 5 of allele *B* result in a threonine to asparagine amino acid change. The third codon change in repeat 7 of allele *B* results in a threonine to alanine change.

with any other sequences in the databases, including the putative signal sequences of *MUC1*, *MUC2* or rat *Muc 2*.

The 5′ flanking sequence (−1945 to +100) of *Mucsmg* was compared with rat submandibular gland-specific genes, including glutamine-glutamic acid rich protein (GRP) variants Ca and Cb [19; Tabak LA, unpublished], rat salivary cystatin S (a cysteine protease inhibitor secreted in post-natal rats but not in adults [20]), the androgen-responsive *VCS-α1* gene [21], and the *VCS-β1* gene [22]. No sequence elements common to all five promotor regions were detected using a window of 10 nt and a 90% homology minimum. This suggests that the *cis*-elements which regulate the tissue-specific expression of these genes are either not located in the sequences immediately flanking the transcription initiation site, or alternatively, that separate mechanisms are responsible for the specificity of their expression. Evidence indicates that greater than 9 kb of flanking sequence is required for the submandibular acinar cell-specific expression of the GRP isoform Ca [23]; however, the level of expression does not approach that which is seen *in vivo*. Smith and coworkers [24] have suggested that submandibular gland expression of the members of the kallikrein gene family is controlled by a dominant locus control region, where

Table 1. Segregation of the rat ***Mucsmg*** gene and chromosome 14 in mouse-rat cell hybrids.

Hybrids	*Rat* Mucsmg *gene*[a]	*Rat chromosomes*[b]																				
		X	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
LB20	−	+	−	(+)	(+)	−	−	(−)	+	−	−	−	−	+	+	−	−	+	(+)	+	+	−
LB150-1	−	+	−	−	+	+	−	−	+	−	+	(+)	+	+	+	−	−	(+)	(+)	+	+	−
LB161	+	+	−	+	+	+	+	+	+	−	+	+	−	(+)	+	+	+	+	+	+	+	(+)
LB210-I	+	+	−	−	−	−	−	−	−	−	−	−	−	−	+	+	−	−	−	+	−	−
LB251	−	+	+	+	−	+	−	(+)	+	−	−	+	−	+	+	−	−	−	+	−	+	−
LB330	−	+	−	+	+	+	−	+	−	−	−	+	−	+	−	−	−	−	+	−	−	−
LB510-6	+	+	−	+	+	+	−	−	+	−	−	−	−	+	+	+	+	+	+	+	−	−
LB600	+	+	+	+	+	+	+	(+)	+	−	(−)	+	+	+	+	+	+	+	−	+	+	−
LB630	+	+	(−)	−	+	+	(+)	+	+	−	+	−	+	+	+	(+)	+	+	−	+	+	(−)
LB780	−	+	−	+	+	+	+	−	+	−	−	+	+	−	+	−	−	−	+	+	−	+
LB810	+	+	−	+	+	+	−	+	+	+	−	+	+	+	+	+	+	+	+	+	−	+	(+)
LB860	−	+	−	+	+	+	−	−	+	−	+	−	+	+	+	−	+	+	+	+	−	(+)
LB1040	−	+	−	−	+	+	(−)	+	+	−	−	+	+	+	−	−	+	+	−	+	−	+
Independent discordant clones[c]*:*		7	5	7	7	7	4	5	7	5	5	8	7	7	5	0	3	5	9	6	5	6

[a]A + or − indicates the presence or absence of the rat gene, respectively.
[b]A + indicates that the rat chromosome is present in more than 55% of the metaphases; (+) indicates that the rat chromsome is present in 25–55% of the metaphases; (−) indicates that the rat chromosome is present in less than 25% of the metaphases; − indicates that the rat chromsome is absent.
[c]Independent hybrid clones are clones derived from distinct fusion events. In this table, all clones are independent. When a chromosome was present in less than 25% of the metaphases (− in parentheses), the hybrid in question was not taken into account to establish the number of discordancies for that particular chromosome.

control of expression may be conferred over extremely large distances.

We have also analysed the *Mucsmg* sequence for elements which have been implicated in regulating the salivary-specific expression of the rat PRP gene, *RP4* [25] and human amylase gene, *AMY1C* [26]. None of these sequence motifs were found in the available 5′ flanking sequence. Salivary gland-specific expression has been achieved with transgenic models of the parotid secretory protein from mouse [27] and with the rat PRP gene, *R15* [28]; however, 11.4 kb and 10 kb, respectively, of the 5′ flanking sequence was required for high levels of tissue-specific expression. Thus, it may be a common theme that high level salivary gland-specific gene expression comes about through the involvement of elements positioned at large distances from the promoter.

To date, only two other mucin genes have been cloned in their entirety: the membrane-bound human tumour-associated polymorphic epithelial mucin, *MUC1*, and the mouse homologue, *Muc 1*. No sequence similarity was found between the promoter sequences of *Mucsmg* and the genes *MUC1* [3] and *Muc 1* [6]. Since these mucins display very different patterns of expression, this is not surprising. In contrast to the other cloned mucins, the RSMG apomucin is encoded by a limited number of exons. The entire secreted portion of the apomucin, as well as the 3′ untranslated sequence is encoded on a single, large exon. Both *MUC1* and its homologue, *Muc 1*, are encoded by seven exons [3, 6]. The finding that the entire secreted RSMG apomucin coding region is contained in a single exon suggests that there may be some selective advantages to maintaining the size of this low molecular weight mucin.

In summary, we have cloned, mapped, and partially sequenced a gene encoding rat submandibular gland apomucin, *Mucsmg*. The 5′ flanking sequence of the *Mucsmg* shows little similarity with any of the cloned salivary-specific genes or other mucin genes. We are currently initiating studies to identify *cis*-elements with the 5′ region of the mucin gene promoter.

Acknowledgments

We thank Christine Cagnina and M. Riviere for their excellent technical assistance. We also thank Ms Patricia Noonan for her help in preparing the manuscript.

This work was supported in part by National Institutes of Health Grant DE08108 (to L.A.T.) and the Fund for Scientific Medical Research (FRSM) and the Belgian program on inter-university attraction poles initiated by the Belgian State Prime Minister's Office (SSTC/DWTC). C.S. is a Research Director of the National Fund for Scientific Research (FNRS).

References

1. Tabak LA (1995) *Ann Rev Physiol* **57**: 547–64.
2. Albone EF, Hagen FK, Van Wuyckhuyse BC, Tabak LA (1994) *J Biol Chem* **269**: 16845–52.
3. Lightenberg MJL, Vos HL, Gennissen AMC, Hilkens J (1990) *J Biol Chem* **265**: 5573–78.
4. Toribara NW, Gum JR Jr, Culhane PJ, Lagace RS, Hicks JW, Petersen GM, Kim YS (1991) *J Clin Invest* **88**: 1005–13.
5. Bobek LA, Tsai H, Biesbrock AR, Levine MJ (1993) *J Biol Chem* **268**: 20563–69.
6. Spicer AP, Parry G, Patton S, Gendler SJ (1991) *J Biol Chem* **266**: 15099–115.
7. Gum JR Jr, Hicks JW, Toribara NW, Sidiki B, Kim YS (1994) *J Biol Chem* **269**: 2440–46.
8. Verma M, Davidson EA (1994) *Glycoconj J* **11**: 172–79.
9. Shirotani K, Taylor-Papadimitriou J, Gendler SJ, Irimura T (1994) *J Biol Chem* **269**: 15030–35.
10. Kovarik A, Peat N, Wilson D, Gendler SJ, Taylor-Papadimitriou J (1993) *J Biol Chem* **268**: 9917–26.
11. Hollingsworth MA, Closken C, Harris A, McDonald CD, Pahwa GS, Maher LJ III (1994) *Nucleic Acids Res* **22**: 1138–46.
12. Chisholm D (1989) *Biotechniques* **7**: 21–23.
13. Sambrook J, Fritsch EF, Maniatis T (1989) In *Molecular Cloning. A Laboratory Manual*, 2nd edition, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
14. Szpirer J, Leven G, Thörn M, Szpirer C (1984) *Cytogenet Cell Genet* **38**: 146–49.
15. Mount SM (1982) *Nucleic Acids Res* **10**: 459–72.
16. McKnight SL, Kingsbury R (1982) *Science* **217**: 316–24.
17. Gum JR Jr (1992) *Am J Respir Cell Mol Biol* **7**: 557–64.
18. Isemura S, Saitoh E (1994) *J Biochem (Tokyo)* **115**: 1101–6.
19. Cooper LF, Tabak LA (1991) *Gene* **106**: 261–66.
20. Cox JL, Shaw PA (1992) *Gene* **110**: 175–80.
21. Rosinski-Chupin I, Rougeon F (1990) *DNA Cell Biol* **9**: 553–59.
22. Courty Y, Rosinski-Chupin I, Rougeon F (1994) *J Biol Chem* **269**: 520–27.
23. O'Connell BC, Ten Hagen KG, Lazowski KW, Tabak LA, Baum BJ (1995) *Am J Physiol* **268**: G1074–78.
24. Smith MS, Lechago J, Wines DR, MacDonald RJ, Hammer RE (1992) *DNA and Cell Biol* **11**: 345–58.
25. Lin HH, Li W, Ann DK (1993) *J Biol Chem* **268**: 10214–20.
26. Ting C, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH (1992) *Genes and Devel* **6**: 1457–65.
27. Larsen HJ, Brodersen CH, Hjorth JP (1994) *Transgen Res* **3**: 311–16.
28. Tu Z, Lazowski KW, Ehlenfeldt RG, Wu G, Lin HH, Kousvelari E, Ann DK (1993) *Gene Expr* **3**: 289–305.