Chapter 7: The major tick salivary gland proteins are part of the tick lipocalin family*

7.1.1 Introduction: The lipocalin protein superfamily

A protein family can be described as proteins of which the amino acid sequence similarity is high enough (>35% identity), to ascribe homology based on sequence alone. In contrast, a protein superfamily consists of proteins with sequence similarity so low (<20% identity) that common origins can only be inferred from structural similarity (Skerra, 2000). Lipocalins form such a superfamily of small (150-183 amino acids residues), extracellular secretory proteins with highly divergent sequences (<20% sequence indentity), but with a highly conserved structural fold (Flower, 1996; Åkerstrom *et al.* 2000; Flower, North and Sansom, 2000). Lipocalins are characterized by their ability to bind small hydrophobic molecules. The name lipocalin was derived from this property: lipo (lipophilic) for their ligands and calyx (ligand enclosed by the protein as is the flower by its calyx) to describe their binding mode (Pervaiz and Brew, 1987).

7.1.2 Lipocalin function

Lipocalins have a diversity of functions that include transport of small molecules, arthropod coloration, pheromone transport, prostaglandin synthesis, smell reception, regulation of cell growth, tissue development and metabolism, regulation of the immune response, allergens and has been implicated in various disease states. Ligands bound by the lipocalins include steroids, retinoids, odorants, pheromones, histamine, haem, nitric oxide and ADP. Lipocalins can also bind to cell surface receptors and can form macromolecular complexes (Flower 1996; Akerstrom *et al.* 2000; Flower, North and Sansom, 2000).

7.1.3 Lipocalin three-dimensional structure

The lipocalins share a similar protein fold consisting of a eight (A-H) stranded, continuous hydrogen-bonded anti-parallel β-barrel with a +1 topology (Fig. 7.1). Sheets are linked by short β-hairpin loops, except loop one that is a large Ω loop that link strands A and B. The structure has a flattened elliptical shape as if composed of two different

orthogonal β-sheets stacked together, with one end closed off by a N-terminal 3_{10} helix. Close packing of the barrel, loops (L2, L4, L6) and helix residues in this area forms a hydrophobic core. The other end is open to solvent and a ligand-binding pocket is formed here via the binding cavity and the exposed loops (L1, L3, L5, L7). A C-terminal α-helix packs against one side of the barrel and is followed by a short β-sheet (I). This topology of a barrel with a central cavity closed at one end and open at the other end combined with the flanking α-helix give the lipocalins the resemblance of a cup. The β-barrel structure is highly conserved for most lipocalins and contributes to the structural stability of the protein fold. In contrast, the loops at the open end are highly variable and contribute to the diversity of ligand and receptor recognition displayed by the lipocalins (Fig. 7.1).

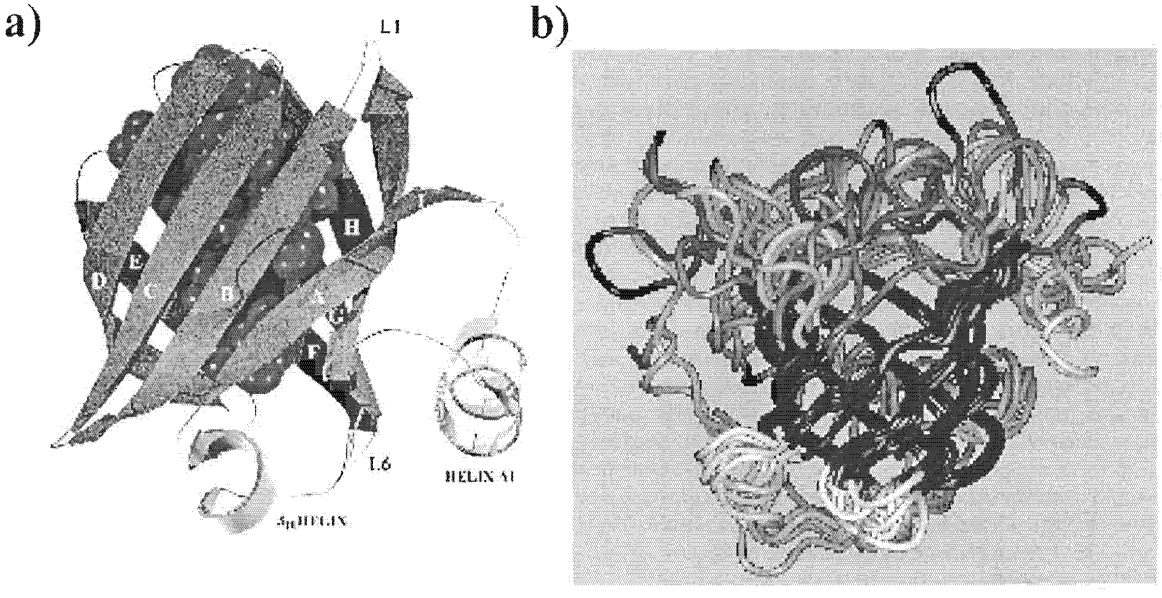


Fig. 7.1: The structure of the lipocalin fold. (a) A schematic drawing of the lipocalin fold with the ligand-binding pocket indicated by space filled spheres. Indicated is the N-terminal 3_{10} helix, the β-sheets (A-H) that form the barrel, the C-terminal α-helix and the single β-strand (I). Adapted from Flower, North and Sansom, (2000). (b) Superposition of six lipocalins that indicates the conserved nature of the β-barrel (black) and closed end (white) and the diversity observed for the variable loops at the open end of the barrel (grey shadings). Adapted from Skerra (2000).

7.1.4 Conserved motifs of the lipocalins

Although highly divergent in sequence, a number of sequence motifs specific for the lipocalins have been described. These structural conserved regions (SCRs) have been used to classify lipocalins into kernel (those that possess all three SCRs) and outlier (those that possess only one or two SCRs). SCR1 is found in the N-terminal 3_{10}-helix and is part of β-strand A. SCR2 comprises regions from both the F and G β-strands, as well as loop L6. SCR3 comprises part of the β-strand H and the C-terminal α-helix, although it corresponds closer to sequence conserved regions than structural topology (Fig. 7.2).

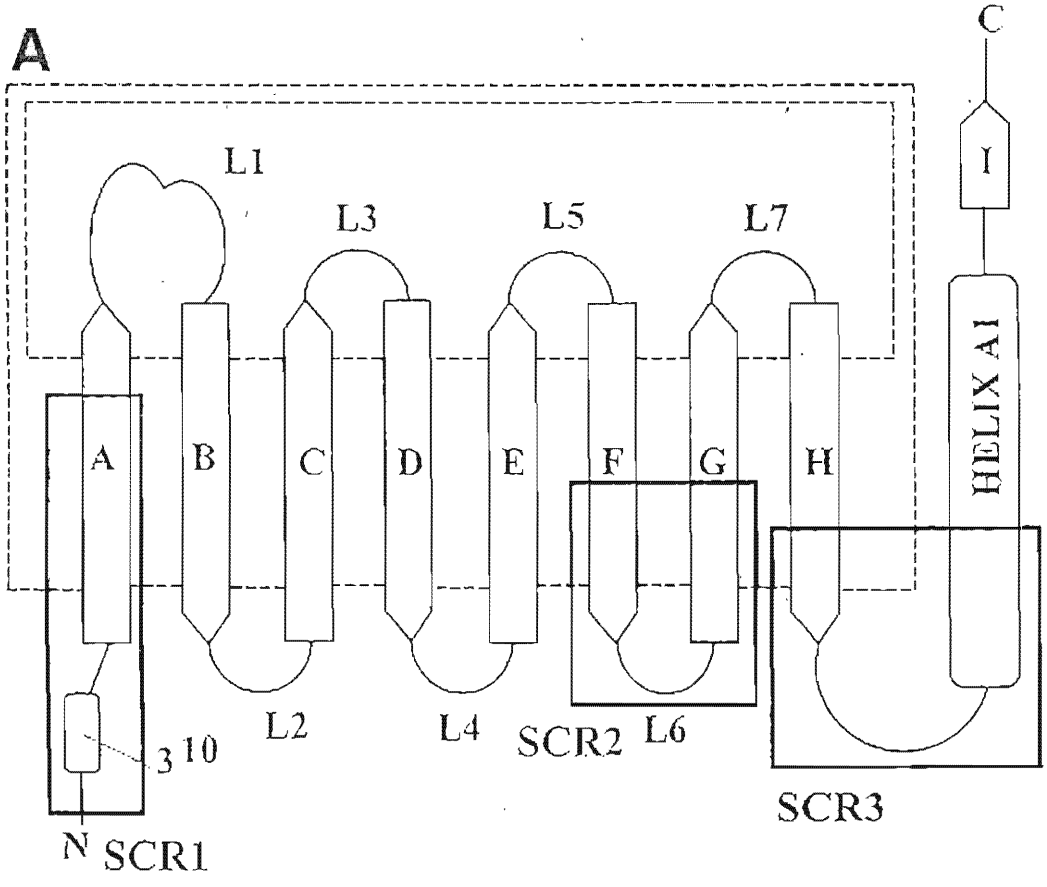


Fig. 7.2: An unwound view of the lipocalin structure. Indicated are the SCRs and their localization, the N-terminal 3_{10}-helix, the +1 topology of the 8 stranded (A-H) β-barrel linked by the short β-hairpin loops (L1-L7), with L2, L4 and L6 at the closed end and L1, L3, L5 and L7 at the open end of the barrel. Adjacent strands are H-bonded indicated by the dotted lines. The C-terminal α-helix and the short β-strand (I) complete the structure. Adapted from Flower, North and Sansom, (2000).

7.1.5 Evolution of the lipocalins

Lipocalins have been identified in prokaryotes, plants and animals (arthropoda and chordata). Prokaryotic and plant lipocalins so far identified are all outlier lipocalins, while both outlier and kernel lipocalins have been identified in animals. This suggests that an outlier lipocalin has been the earliest ancestor and that the conserved motifs of the kernel lipocalins have only appeared at a step in animal evolution that preceded arthopod/chordata divergence. All kernel lipocalins were probably derived from a single ancestor and could probably be traced more easily using phylogenetic methods than the outlier lipocalins that are much more diverged (Salier, 2000).

Evolution of lipocalins has been described using sequence distance, maximum parsimony and maximum likelihood methods. All methods group the lipocalins into various groups according to conserved function, although distance and parsimony methods failed to make any specific conclusions about the relationship between the functional groups (Ganfornina *et al.* 2000; Gutierrez, Ganfornina and Sanchez, 2000). Maximum likelihood was useful to describe relationships between functional groups so that a lineage could be traced from prokaryotic, through arthopoda up to the more recent mammalian lipocalins (Fig. 7.3). The mammalian lipocalins can be divided into ancient and modern lipocalins based on taxonomic representation. The lipocalins from clade II are considered to be ancient metazoan lipocalins based on their presence in arthropods and chordates. Clades III, V and VI are considered to be ancient vertebrate lipocalins and those from IV, VII-XIV to be modern lipocalins. General trends deduced from this analysis included a higher rate of sequence divergence and gene duplication, a reduction in binding surface area and an increase in ligand-binding contacts for modern lipocalins (Ganfornina *et al.* 2000; Gutierrez, Ganfornina and Sanchez, 2000). This analysis was biased however, in that only sequences that show more than 20% sequence identity to its closest relatives and sequences that contained at least two SCR motifs were used for this study. This means that many of the outlier lipocalins were left out from the study, which in evolutionary terms probably show the highest divergence of all lipocalins. It was stated that the outlier lipocalins were not included in the phylogenetic analysis due to problems with sequence alignment.

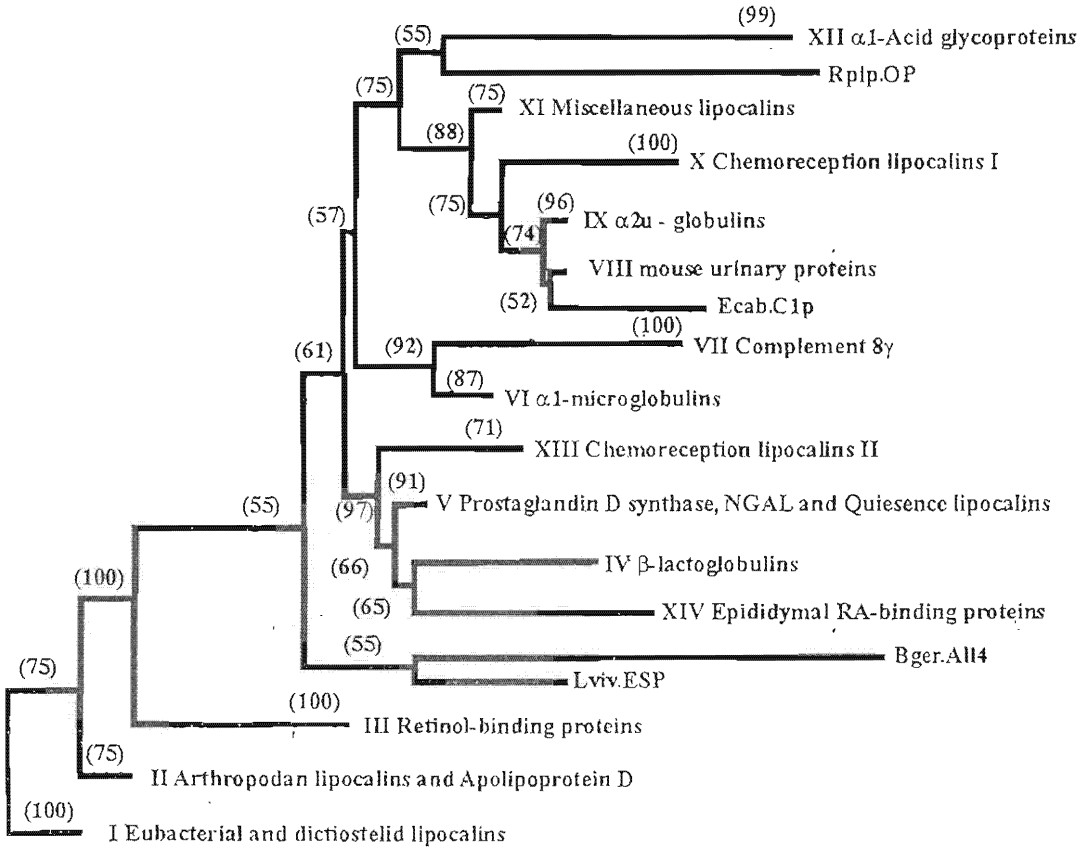


Fig. 7.3: A phylogenetic tree of the lipocalins. Eubacterial and dictyostelid lipocalins were used as outgroup (Clade I), followed by arthopod and mammalian apolipoprotein D lipocalins (clade II), found in metazoans. This is followed by the retinal binding proteins (clade III), β-lactoglobulins (clade IV) and prostaglandin D synthases (clade V) found only in vertebrates. Lipocalins found only in mammals give a more complex relationship, in that some group with lipocalins found only in vertebrates (the chemoreception lipocalins, clade XIII and epididymal RA-binding proteins, clade XIV) and other groups in in higher clades (VI-VIII, IX-XII). Adapted from Gutierrez, Ganfornina and Sanchez (2000).

7.1.6 Lipocalins from hematophagous organisms

Lipocalins have been found in the salivary gland secretions from various hematophagous organisms. Lipocalins identified include seven from *Rhodnius prolixus*, two from *Triatoma pallidipennis*, one from the hard tick *R .appendiculatus*, at least two from the soft tick *O. moubata* and at least four from the soft tick *O. savignyi* (Montford, Weichsel and Andersen, 2000; Mans *et al.* 2001). In *R. prolixus* at least four lipocalins carry heme and are called the nitrophorins (NP1-NP4). All four proteins carry nitric oxide, which is released at the feeding site, causing smooth muscle relaxation and vasodilation (Champagne, Nussenzvieg and Ribeiro, 1995). All four proteins can also regulate

inflammation and the host's immune response by binding histamine. NP2 (prolixin-S) can also inhibit the conversion of fX to fXa (Ribeiro, Schneider and Guimares, 1995). Another three lipocalins from *R. prolixus* do not contain heme, but inhibit platelet aggregation by sequestration of the agonist ADP (Francischetti *et al.* 2000). *T. pallidipennis* inhibits specifically collagen-induced platelet aggregation via the lipocalin, pallidipin (Noeske-Jungblut *et al.* 1994). Triabin is another lipocalin that inhibits thrombin's activity by targeting the fibrinogen-binding exosite (Noeske-Jungblut *et al.* 1995). In the case of triabin an interesting deviation in its topology occurred with an exchange of β-strands B and C of the β-barrel (Feuntes-Prior *et al.* 1997).

In hard ticks, the histamine-binding proteins (HBP1-3) have been identified in *R. appendiculatus* (Paesan *et al.* 1999). Two histamine-binding sites have been described in these lipocalins and they presumably function in the regulation of inflammation during the long feeding periods of hard ticks (Paesan *et al.* 2000). The structurally conserved motifs (SCR1-3) used for lipocalin designation are present in the core lipocalins, while one or more are normally absent in the outlier lipocalins (Flower, 1996; Flower *et al.* 2000). All SCRs are absent in the tick lipocalins although the structure of HBP2 shows lipocalin topology (Paesen *et al.* 1999). Moubatin, an inhibitor of collagen-stimulated platelet aggregation has been described for the soft tick, *O. moubata* and is also a lipocalin based on sequence similarity to the HBPs (Waxman and Connolly, 1993; Keller *et al.* 1993; Paesen *et al.* 1999). TAI an inhibitor of collagen specific cell adhesion (Karczewski *et al.* 1995) has also been identified as a lipocalin based on a sequence contained in a patent (European patent application number 92311218.9) (Guido Paesen and Patricia Nuttall, personal communication).

7.2 Materials and methods

7.2.1 Cloning and sequencing of TSGPs

The strategy used to clone and sequence the TSGPs were described in Chapter 2. Degenerate primers were designed from the obtained N-terminal amino acid sequences. For TSGP1 a primer (20kD: GGI CCI GAY GGI TGY GT) was designed using the first 6 amino acids (GPDGCV), for TSGP2 the primer (TOE: TTY CCI ACI GAR GCN TA) was

designed from amino acids 6-11 (FPTEAY) and for TSGP3 the primer (TOC: TTY CCI ACI GAY GCN TA) was also designed from amino acids 6-11 (FPTDAY), while the primer for TSGP4 (Toks1: GCN AAY GAY GTI TGG AAY GT) was designed from the first 7 amino acids (ANDVWNV). Note that amino acid residue 6 of both TSGP2 and TSGP3 are indicated as phenylalanine. Results obtained with 3'RACE indicated that this position in both sequences is a lysine. This was due to an initial misidentification of lysine for phenylalanine during N-terminal sequence analysis. Original primers (TOKSA: GAY TGY CCN ACN GGI TTY C; TOKSB: GAY TGY CCN ACN GGI TTY CCI AC) designed from the first 6 or 8 amino acids (DCPTGF, DCPTGFPT) yielded spurious products due to mispriming of the phenylalanine. TOKSA amplified ferritin (a 600 bp product), while TOKSB amplified actin (800 bp product) (results not shown). 3'RACE and 5'RACE were performed as described using an annealing temperature of 55 °C and performing 27 cycles. 5' RACE primers used were for TSGP1 (20KDC1: GTG TAG GGG ATG GGG CCA), TSGP2 and (C1T2: CTA GCA GTC CTT GTC TT) and TSGP3 (NTC1: GTT CCA ACA TCC ACA TG), TSGP4 (C1T1: CTA CGG AAC TCT GCA GCC TT). 20KDC1 is complementary to a region in the 3' untranslated region of TSGP1. C1T2 are complementary to the last five amino acids and stop codon of TSGP2 (QDKDC-). NTC1 is complementary to an internal sequence (DMWMLE) of TSGP3 that differs from that of TSGP2 (EMWMLE) at the last position in the codon of aspartic acid. C1T1 is complementary to the last five amino acids and stop codon of TSGP4 (EGCRVP-). Sequences were analyzed as described in Chapter 2.

7.2.2 Multiple alignment of lipocalins

Multiple alignment was performed using ClustalX (Jeanmougin *et al.* 1998), using an identity matrix based on the secondary structures of crystallized lipocalins (Ganfornina *et al.* 2000; Gutiérrez, Ganfornina and Sanchez, 2000).

7.2.3 Phylogenetic analysis of lipocalins

Phylogenetic analysis was performed on sequences in which all gapped positions were ignored using Neighbor-Joining analysis with the Mega2 software package (Kumar, Tamura and Nei, 1994).

7.2.4 Molecular modeling of lipocalins

Structures of the TSGPs were modeled on that of Ra-HBP2 (Paesen *et al.* 1999) using MODELLER (Sali *et al.* 1995). Structures were validated using PROCHECK (Laskowski *et al.* 1996) and WHATIF (Vriend, 1990) and backbone deviations obtained using ProFitV1.8 (http://www.biochem.ucl.ac.uk/~martin/#profit). Surface models were generated with GRASP (Nicholls, Sharp and Honig, 1991).

7.2.5 Partial purification of savignygen

During the present study it became clear that TSGP2 and TSGP3 showed high sequence similarity (73%) to moubatin (Keller *et al.* 1993). In order to investigate the possibility that TSGP2 or TSGP3 might be moubatin orthologs, SGE was fractionated and assayed for inhibitory activity towards collagen-induced platelet aggregation. The microplate method described in Chapter 2 was used, but instead of activating platelets with ADP, 2μg collagen (Diagnostica Stago) was used. Before use collagen (100 μl of 2 mg/ml stock solution) was first diluted in 1900 μl 150 mM Tris-HCl, pH 7.8 and incubated at 37 °C for 3 minutes. SGE was fractionated using RPHPLC as described in Chapter 6 and all fractions assayed by drying 100 μl (from each milliliter fraction) and redissolving in 100 μl buffer (0.15M NaCl, 20 mM Tris-HCl, pH 7.4) before using 10 μl for each experiment. All experiments were done in duplicate. Positive fractions were then dried before application to AEHPLC as described in Chapter 2. Fractions collected were again tested before rechromatography using RPHPLC. As sequence similarity between TSGP2 and TSGP3 have been indicated, fractions from the RPHPLC were also probed using western blotting with polyclonal sera directed against TSGP2.

7.3 Results

7.3.1 RACE of the TSGPs

RACE procedures were performed for the TSGPs. 5'-RACE and 3'-RACE under optimized conditions gave products for TSGP1 (~600bp, ~600bp), TSGP2 (~600bp, ~500bp), TSGP3 (~600bp, ~500bp), TSGP4 (~600bp, ~650bp), respectively (Fig. 7.4). The 5'-RACE products include the 5'UTR, the gene coding for the mature as well as immature protein. The 3'-RACE products include the coding gene, 3'-UTR and poly-A tail sequences.

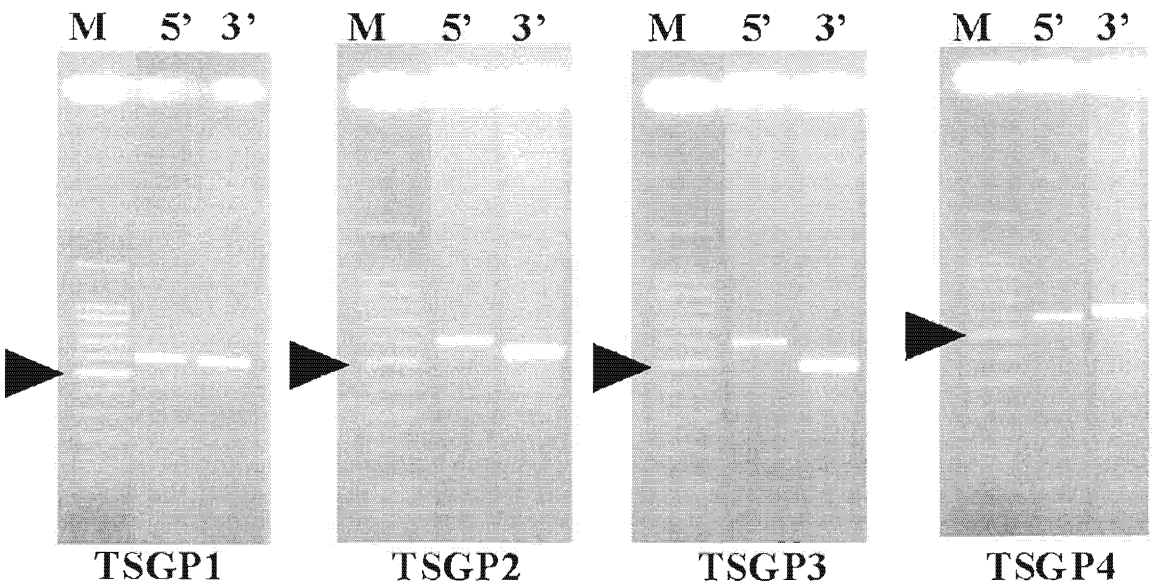


Fig. 7.4: RACE of the TSGPs. Indicated are the results for 5'-RACE and 3'-RACE of TSGP1-4. Arrowheads indicate the 500 bp band.

7.3.2 Amino acid sequences of the TSGPs

Full-length sequences for the TSGP1-4 and savignygen were obtained by overlapping of 5'RACE and 3'RACE product sequences. The sequence lengths corresponded well with the lengths of the RACE products obtained (results not shown). The full-length gene sequences of TSGP1 (673bp), TSGP2 (609bp) and TSGP3 (610bp) and TSGP4 (673bp) all include a stop codon, poly-adenylation site (TSGP1/TSGP4: AATAAA and TSGP2/TSGP3: AGTAAA) and a poly-A tail (Fig. 7.5-7.7). The AGTAAA site was previously also identified in savignin, a thrombin inhibitor from this tick species (Chapter

3). The translated amino acid sequences of the immature proteins contain a signal peptide and consisted of 190 amino acids (TSGP1), 163 amino acids (TSGP2 and TSGP3) and 176 amino acids (TSGP4). Signal P predicted the presence of the signal peptide and the correct cleavage site in all cases (von Heijne, 1990; Nielsen *et al.* 1997). The mature proteins consist of 171 amino acids (TSGP1), 144 amino acids (TSGP2 and TSGP3) and 156 amino acids (TSGP4) that include the previously determined N-terminal sequences (Chapter 6). Of interest is Glu16 in the mature TSGP2/TSGP3 sequences, which showed up as a unidentified amino acid during N-terminal sequencing (Chapter 6). The elution profile of this residue during N-terminal sequencing (personal observation), possibly indicate carboxyl methylation, which could probably be involved in salivary gland granule packaging and exocytotic secretion (Van Waarde, 1987).

```
tcactatagggctcgagcggccccgcccgggcaggtaagatgcaacggc                    M  Q  R  L  L  L  L      7'

ctgattgccctgttctcgctcagctgtgctgaagcagggccggatgggtgcgtgggtagt  -  120
 L  I  A  L  F  S  L  S  C  A  E  A  G  P  D  G  C  V  G  S      27'

acagaggctaaggtggctgtatttggagaaggtggaaatgcaggatctccaactataggg  -  180
 T  E  A  K  V  A  V  F  G  E  G  G  N  A  G  S  P  T  I  G      47'

tactcttaccttgtgaagacaacctatcctgatgaacatgcttgtgtttacattcttcca  -  240
 Y  S  Y  L  V  K  T  T  Y  P  D  E  H  A  C  V  Y  I  L  P      67'

ccctatggcacagcggacgctagtggccgctacccttaccgcatggggtacaaggactca  -  300
 P  Y  G  T  A  D  A  S  G  R  Y  P  Y  R  M  G  Y  K  D  S      87'

aacgatcagtgggtgaagctggatgggaagatcaaaaccgagggcagcaaaatcatcgac  -  360
 N  D  Q  W  V  K  L  D  G  K  I  K  T  E  G  S  K  I  I  D     107'

aacgacccggaatatggcgacactgtgaccacggtgctctacactcaccttgggggtgga  -  420
 N  D  P  E  Y  G  D  T  V  T  T  V  L  Y  T  H  L  G  G  G     127'

tgtgacgttacactcttcgaagggcaaaagggccagagcaaagtacaaggaccattcctg  -  480
 C  D  V  T  L  F  E  G  Q  K  G  Q  S  K  V  Q  G  P  F  L     147'

gaactgtggtaccacagtggagcaagtgaagaatccatgcggtgctgcgaggaagagttt  -  540
 E  L  W  Y  H  S  G  A  S  E  E  S  M  R  C  C  E  E  E  F     167'

aggaagaatcttaaggaagggacggctgttcgaaaggttaacaagaactgtgactatggg  -  600
 R  K  N  L  K  E  G  T  A  V  R  K  V  N  K  N  C  D  Y  G     187'

gacgtcgcctagaagaatgcgtgaactggccccatcccctacacgaacgctcgaaaaaaaa  -  660
 D  V  A  -                                                     190'

aaaaaataaagaaagaaatacaaatcataaaaaaaaaaaaaaaaaaaagagtgttgtggt  -  720

aatgatagc                                                      -  729
```

Fig. 7.5: Full-length sequence of TSGP1. The 5' adapter, 3' gene specific and 3' anchor primers are shown in bold. The stop codon (TAG), poly-adenylation site (AATAAA) and poly-A tail are boxed. The N-terminal amino acid sequence previously obtained with N-terminal Edman degradation is underlined in a solid line and the signal sequence is underlined with a dashed-line. The N-terminal sequence used for degenerate primer design is shown in bold. The Genbank accession code for TSGP1 is AF452888.

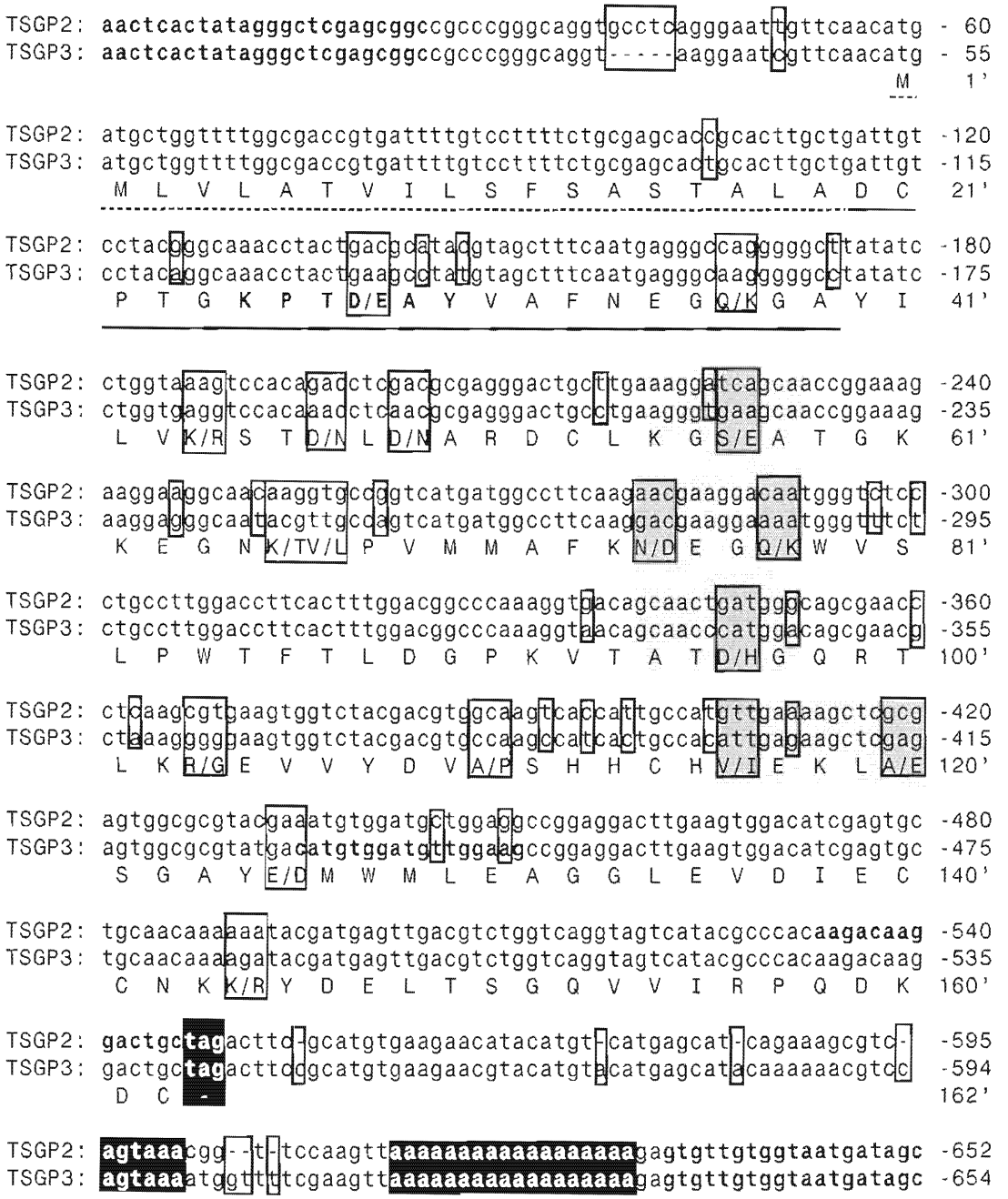


Fig. 7.6: Full-length sequences of TSGP2 and TSGP3. Synonymous nucleotide differences are boxed, while non-synonymous differences are boxed in gray. The 5' adapter, 3' gene specific and 3' anchor primers are shown in bold. The stop codon (TAG), poly-adenylation site (AGTAAA) and poly-A tail are indicated by black boxes. For the deduced amino acid sequences, the differences are indicated by a slash (TSGP2/TSGP3). The N-terminal amino acid sequences previously obtained with N-terminal Edman degradation is underlined in a solid line and the signal sequence is underlined with a dashed-line. The N-terminal sequences used for degenerate primer design is shown in bold. The Genbank accession code for TSGP2 and TSGP3 are AF452889 and AF45890, respectively.

```
aactcactatagggctcgagcggcccgcccgggcaggtcgataaaca

tcgctgcgtaacggaacgaatatggactgcaagcttgtcgccatcgcgctcttcattttc - 120
                          M  D  C  K  L  V  A  I  A  L  F  I  F    13'
tccctagattttgcacatgcggctaacgacgtatggaacgtcctcaaaggcagcgattca - 180
 S  L  D  F  A  H  A  A  N  D  V  W  N  V  L  K  G  S  D  S    33'

aagtttcttatggtcaagagaacatatgaaagaggagcaaacaaatgtgtgtacatgaaa - 240
 K  F  L  M  V  K  R  T  Y  E  R  G  A  N  K  C  V  Y  M  K    53'

cgtacgagcatggacgaaagcagtcatacacttgaagtacttatgggatattcgaaggcg - 300
 R  T  S  M  D  E  S  S  H  T  L  E  V  L  M  G  Y  S  K  A    73'

gggacaacgacggacttcgtagagccatctaagtatactgtgacagcaactagtgagggt - 360
 G  T  T  T  D  F  V  E  P  S  K  Y  T  V  T  A  T  S  E  G    93'

gcaagcacctacaatatgatgactgtgagaaggggacctgcctcgcatggtgtcaaattc - 420
 A  S  T  Y  N  M  M  T  V  R  R  G  P  A  S  H  G  V  K  F   113'

gagctggtgtacagcgatgaccaaggctgcaatattctgcaaatgaagacgagtccattt - 480
 E  L  V  Y  S  D  D  Q  G  C  N  I  L  Q  M  K  T  S  P  F   133'

ccaggaaaatgcgaactgtgggcgccggaaggcaaggcaaagaatgtggaaagcagttgc - 540
 P  G  K  C  E  L  W  A  P  E  G  K  A  K  N  V  E  S  S  C   153'

agcggcaagttcaaggagttatgtggcgacgcagtggaaacgccctacgcagaaggctgc - 600
 S  G  K  F  K  E  L  C  G  D  A  V  E  T  P  Y  A  E  G  C   173'

agagttccgtagttcctggccagattcacagttgtggaactgttttctgaacagactaat - 660
 R  V  P  -                                                    176'

aaagctttaaggcagacagagcaaaaaaaaaaaaaaaaaagagtgttgtggtaatgatag - 720
```

Fig. 7.7: Full-length sequence of TSGP4. The 5' adapter, 3' gene specific and 3' anchor primers are shown in bold. The stop codon (TAG), poly-adenylation site (AATAAA) and poly-A tail are boxed. The N-terminal amino acid sequence previously obtained with N-terminal Edman degradation is underlined in a solid line and the signal sequence is underlined with a dashed-line. The N-terminal sequence used for degenerate primer design is shown in bold. The Genbank accession code for TSGP4 is AF452891.

7.3.3 Comparison of data from native toxins and deduced amino acid sequences

Amino acid compositions of the deduced amino acid sequences compare favorably with data from the native proteins (Table 7.1). The calculated molecular masses also correlate well with those obtained from the ESMS analysis. From these data it is clear that the toxins are not glycosylated, in contrast to previous reports (Neitz *et al.* 1983).

Table. 7.1: Comparison of the amino acid composition from the TSGPs and their deduced amino acid sequences. Indicated are molar ratio's determined using Ile as 1. In the case of TSGP1, TSGP2 and TSGP3 the values were multiplied by a factor that gives six cysteines, as this was determined independently using MALDI-TOF-MS. Also indicated are calculated molecular masses from the amino acid composition and deduced sequences as well as masses obtained from ESMS.

Amino acids	TSGP1		TSGP2		TSGP3		TSGP4	
	AA	Seq	AA	Seq	AA	Seq	AA	Seq
Asx	15	18	15	16	14	15	11	13
Glx	16	17	16	15	16	15	14	14
Ser	10	10	6	6	5	5	14	15
Gly	21	23	13	13	14	14	13	13
His	3	3	3	3	4	4	2	2
Arg	3	5	4	4	4	4	6	6
Thr	11	12	11	11	12	12	13	13
Ala	10	9	10	10	8	8	9	10
Pro	8	8	6	6	7	7	7	7
Tyr	9	12	5	5	5	5	6	7
Val	13	13	12	13	9	11	12	13
Met	1	2	4	4	4	4	6	7
Cys	6	6	6	6	6	6	6	6
Ile	5	5	3	3	4	4	1	1
Leu	9	9	10	10	11	11	8	8
Phe	4	4	3	3	3	4	5	5
Lys	13	13	13	13	13	12	15	14
Total	157	169	140	141	139	141	148	154
Mr	16699	18613	15238	15872	15328	15950	16143	17161
ESMS		18422		15877		15957		17170

Theoretical trypsin peptide mapping corresponds well with peptide mass fingerprints previously obtained (Table 7.2). It also shows correspondence between peptide fragments across the full-length of the TSGP sequences. These results indicate that the correct sequences have been obtained. In the case of TSGP1 there were however, a few peptides that could not be matched to the sequence. RPHPLC did however, indicate that TSGP1 eluted as a broad tailing peak (Chapter 6) that could indicate microheterogeneity on sequence level.

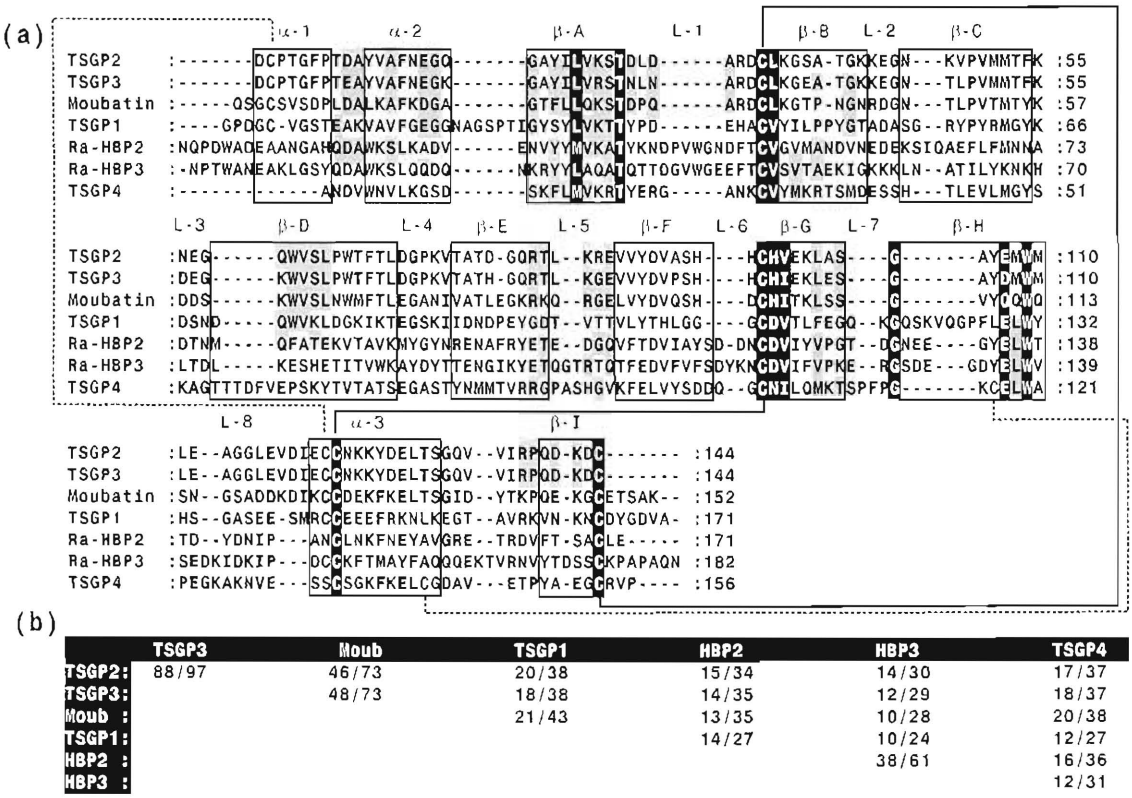
Table 7.2: A comparison of a theoretical tryptic digest of the TSGPs with peptide mass fingerprints previously obtained by MALDI-TOF-MS. Amino acid residues for the different fragments are indicated in parenthesis.

TSGP1		TSGP2		TSGP3		TSGP4	
MS	Sequence	MS	Sequence	MS	Sequence	MS	Sequence
1226	1-12 (1225)			869.8	869 (74-81)	2173.8	2171.4 (16-34)
599.2	59-62 (597.7)	1858.1	1858 (86-100)	1954.5	1954 (85-100)	2085.8	2083.3 (65-83)
2691	143-163 (2697)	1977.5	1977 (126-142)	1976.7	1976 (126-142)	909	908 (84-93)
1319		1849.4	1848 (127-142)	1848.2	1848 (127-142)	2110	2108.4 (93-109)
1391						1138.7	1137.3 (117-125)
1754						1925.7	1923.2 (139-154)

7.3.4 Multiple alignment of tick lipocalins

BLAST analysis of the TSGPs indicated identity to moubatin, a collagen-specific platelet aggregation inhibitor (Fig. 7.8) (Keller *et al.* 1993). Moubatin shows distant similarity to the histamine-binding proteins from the tick *R. appendiculatus* (Paesen *et al.* 1999). Alignment of TSGP1-4 with moubatin and the female HBP2 and male HBP3 shows that there is significant similarity, although identities are very low between sequences. The structural conserved motifs (SCR1-3) used for lipocalin designation are present in the core lipocalins, while one or more are normally absent in the outlier lipocalins (Flower, 1996; Flower *et al.* 2000). All SCRs are absent in the tick lipocalins although the structure of HBP2 show lipocalin topology (Paesen *et al.* 1999). The highest conserved

regions of the tick lipocalins correspond with that of the secondary structure previously obtained for HBP2 and consists of two N-terminal α-helixes, a 8 stranded anti-parallel β-barrel with a $(+1)_7$ topology and a C-terminal α-helix, characteristic of the lipocalin fold (Paesen *et al.* 1999).



	TSGP3	Moub	TSGP1	HBP2	HBP3	TSGP4
TSGP2:	88/97	46/73	20/38	15/34	14/30	17/37
TSGP3:		48/73	18/38	14/35	12/29	18/37
Moub :			21/43	13/35	10/28	20/38
TSGP1:				14/27	10/24	12/27
HBP2 :					38/61	16/36
HBP3 :						12/31

Fig. 7.8: Multiple sequence alignment of the tick lipocalins. (a) Alignment of TSGPs with the HBPs from the hard tick, *R. appendiculatus* and moubatin, from the soft tick *O. moubata*, the inhibitor specific for collagen-induced platelet aggregation. Secondary structures based on that of Ra-HBP2 are boxed and designated as α-helixes or β-strands. Solid lines indicate conserved cysteines and their corresponding disulphide bonds, as deduced from the structure of Ra-HBP2. Dotted lines indicate hypothetical disulphide bonds of the remaining cysteines for moubatin, TSGP1-3 and TSGP4. (b) Percentage identity/similarity between the different sequences are indicated.

7.3.5 Phylogenetic analysis of tick derived lipocalins in relation to the lipocalin family

Previous phylogenetic analysis of the lipocalins excluded those from blood-feeding organisms, due to the low sequence similarity (<20% identity) with other lipocalins and the absence of SCR motifs (Ganfornina *et al.* 2000; Gutiérrez *et al.* 2000). The extreme

divergence of these lipocalins can introduce serious artefacts in the phylogenetic trees due to long branch attraction. While it would thus be difficult to determine an accurate relationship of tick lipocalins within the larger lipocalin family, phylogenetic analysis could be used to assess their homology and investigate their relationships within a tick lipocalin clade. The alignment of the lipocalin family previously employed to investigate lipocalin evolution (Ganfornina *et al.* 2000; Gutiérrez *et al.* 2000), was used as a profile to align both tick lipocalins as well as lipocalins from triatomine bugs. It is clear that the lipocalin family is highly divergent as exemplified by the low levels of sequence similarity indicated (Fig. 7.9). Only a few residue sites are conserved across the family and correspond to the SCR regions as indicated. From this it is clear that tick lipocalins are outliers even though a few residues in the SCRs are also found in the tick lipocalins.

Figures on following pages

Fig. 7.9: Alignment of the lipocalin family used for phylogenetic analysis. Indicated are the different monophyletic clades into which the lipocalins are grouped as well as the regions corresponding to SCRs of core lipocalins. Indicated are similarities based on the PAM 250 matrix (DNQH, SAT, KR, FY, LIMV) at 80% identity.