J Mol Evol (1990) 31:40–50

Journal of
Molecular Evolution

A New Member of a Secretory Protein Gene Family in the Dipteran *Chironomus tentans* Has a Variant Repeat Structure

Joakim Galli,[1] Urban Lendahl,[1] Gabrielle Paulsson,[1] Christer Ericsson,[1]* Tomas Bergman,[2] Mats Carlquist,[3]† and Lars Wieslander[1]

[1] Department of Molecular Genetics, Medical Nobel Institute, Karolinska Institutet, Box 60400, S-104 01 Stockholm, Sweden
[2] Department of Physiological Chemistry, Karolinska Institutet, Box 60400, S-104 01 Stockholm, Sweden
[3] Department of Biochemistry II, Karolinska Institutet, Box 60400, S-104 01 Stockholm, Sweden

Summary. We describe the structure of a gene expressed in the salivary gland cells of the dipteran *Chironomus tentans* and show that it encodes 1 of the approximately 15 secretory proteins exported by the gland cells. This sp115,140 gene consists of approximately 65 copies of a 42-bp sequence in a central uninterrupted core block, surrounded by short nonrepetitive regions. The repeats within the gene are highly similar to each other, but divergent repeats are present in a pattern which suggests that the repeat structure has been remodeled during evolution. The 42-bp repeat in the gene is a simple variant of the more complex repeat unit present in the Balbiani ring genes, encoding four of the other secretory proteins. The structure of the sp115,140 gene suggests that related repeat structures have evolved from a common origin and resulted in the set of genes whose secretory proteins interact in the assembly of the secreted protein fibers.

Key words: Gene family — Repetitive sequences — Structural proteins — Balbiani ring genes

Introduction

The salivary gland cells in the dipteran *Chironomus tentans* synthesize and export approximately 15 different secretory proteins. These proteins are stored in the gland lumen and, upon excretion through the excretory duct, assemble into water-insoluble fibers, which then are spun into a network forming the larval feeding and housing tube (Grossbach 1977).

The secretory proteins thus take part in a multistep process in which their interactions have to be controlled to achieve the final common function, to build the larval tube. The corresponding set of genes should therefore have evolved in an interdependent fashion. In order to better understand the evolutionary pathways and the underlying mechanisms that have resulted in the set of cooperating genes, we wish to characterize the involved genes and establish if and how they are related to each other.

The major protein components in the secretion of the gland cells are the four huge sp-I proteins, each with a relative molecular mass of about 10^3 kd (Edström et al. 1980; Rydlander and Edström 1980; Kao and Case 1985), encoded in the Balbiani ring (BR) genes, BR1, BR2.1, BR2.2, and BR6 (Case 1986; Botella et al. 1988). These genes all have the same type of internal repetitive structure (for references see Pustell et al. 1984; Wieslander et al. 1984; Grond et al. 1987).

In each gene, more than 100 almost identical repeats are tandemly organized in a core block. One half of each repeat unit, the subrepeat (SR-) region consists of short subrepeats, and the other half, the constant (C-) region, contains four conserved cysteine codons. This complex repeat structure with alternating SR- and C-regions has probably arisen from a simple repeat structure in an ancestor com-

* *Present address:* Molecular Biology and Virology Laboratory, The Salk Institute, PO Box 85800, San Diego, California 92138 USA
† *Present address:* Karo Bio, Box 4032, 141 04 Huddinge, Sweden
Offprint requests to: L. Wieslander

mon to all four BR genes via several intermediate stages of gradually more complex repeat units (Pustell et al. 1984; Höög et al. 1988).

Here we isolate and characterize a gene related to the BR genes, the sp115,140 gene. This gene is located in a separate chromosomal locus and encodes a secretory protein with a relative molecular mass of 115 kd or 140 kd. The gene has a highly repetitive structure, in which 65–68 copies of a 42-bp sequence are arranged in tandem in a core block, surrounded by short nonrepetitive regions. The 42-bp repeat sequence is very similar and in large parts identical to the subrepeat sequences in the SR-regions of the BR genes. The sp115,140 gene therefore has a simple repeat unit that corresponds to only one subrepeat of a BR gene SR-region. We discuss evolutionary and functional implications of the determined gene structure.

Materials and Methods

Extraction of DNA, RNA, and Proteins. High molecular weight DNA was extracted from cultured *C. tentans* epithelial cells (Wyss 1982) as described (Gross-Bellard et al. 1973).

For extraction of RNA and proteins, salivary glands were dissected manually from fourth instar larvae, fixed in 70% ethanol at 4°C and stored in glycerol : ethanol (1:1) at −20°C. RNA was extracted from the fixed glands in 1 mM EDTA, 20 mM Tris-HCl, pH 7.4, containing 0.5% SDS and 0.1 mg/ml proteinase-K for 30 min at room temperature, followed by phenol extraction and ethanol precipitation. RNA to be used as template for sequencing was pelleted through a cushion of 5.7 M CsCl in 0.1 M EDTA, pH 7.5, in an SW 50 rotor at 35,000 rpm for 12 h.

For extraction of gland lumen proteins, all cells were removed from the fixed glands with dissection needles. Proteins were extracted in 6 M guanidine hydrochloride, reduced in 0.14 M mercaptoethanol, and incubated for 1 h at room temperature in the presence of 0.1 M iodoacetamide according to Kao and Case (1985).

Construction and Screening of Complementary DNA (cDNA) and Genomic Libraries. Poly(A)$^+$ RNA was obtained by oligo(dT)-cellulose chromatography of total gland RNA extracted as above. The RNA was reverse transcribed using oligo(dT) priming and made double stranded essentially according to Gubler and Hoffman (1983) as modified by Amersham. EcoRI linkers were added, and the cDNA was ligated to λgt10 vector arms (Promega). Labeled single-stranded cDNA was transcribed from poly(A)$^+$ RNA using ^{32}P-labeled precursors. ^{32}P-labeled probes, representing the repeat units of the BR1, BR2.1, BR2.2, and BR6 genes were obtained by SP6 and T7 RNA polymerase transcription of subcloned parts of the repeat units as described (Lendahl and Wieslander 1987).

A *C. tentans* partial Sau3A genomic library in the EMBL4 vector (Paulsson et al. 1989) was screened with a subcloned DNA fragment containing six repeat units of the 6:22 cDNA, labeled by random priming (Feinberg and Vogelstein 1983). Plaque hybridizations were performed according to Benton and Davis (1977). Phage DNA was prepared from liquid cultures (Maniatis et al. 1982).

In Situ Hybridization. The 6:22 cDNA insert was cleaved out with EcoRI, eluted from agarose gels using DEAE membranes (Dretzen et al. 1981), and subcloned into the Bluescript vector (Stratagene). ^{3}H-labeled RNA transcripts were obtained by T7 RNA polymerase transcription in the presence of ^{3}H-UTP and ^{3}H-CTP (New England Nuclear). Salivary gland squash preparations were obtained and the ^{3}H-labeled RNA hybridized as previously described (Sümegi et al. 1982). Autoradiography was performed with Ilford K-2 liquid emulsion.

Northern, Southern, and Western Blots. Total salivary gland RNA was denatured for 5 min at 60°C in 40 mM triethanolamine-HCl, pH 7.4, 10 mM EDTA, containing 50% formamide and 2.2 M formaldehyde. The RNA was then electrophoresed in formaldehyde-containing agarose gels as described (Lehrach et al. 1977), except that the gel buffer and electrophoresis buffer consisted of 40 mM triethanolamine-HCl, pH 7.4, 10 mM EDTA. The RNA was transferred to nylon filters (Amersham). As size markers, denatured DNA fragments of known lengths were run in parallel and detected by hybridization with ^{32}P-labeled DNA of the same kind. Southern blots were performed as described (Southern 1975).

Hybridization of labeled oligodeoxynucleotides was carried out as described (Wallace and Miyada 1987). In vitro RNA transcripts and subcloned DNA fragments were hybridized according to Thomas (1980).

For immunological detection of size-separated proteins, the reduced and alkylated secretory proteins were precipitated in acetone and electrophoresed in SDS-containing polyacrylamide gels (Laemmli 1970). The gels were 3–20% concave exponential gradient gels made according to Kao and Case (1985). The proteins were transferred to nitrocellulose filters as described (Burnette 1981), except that the electrode solution was supplemented with 0.1% SDS. The dried filters were incubated in 10 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.05% Tween-20, 1% bovine serum albumin to block unspecific binding sites, followed by incubation with affinity-purified oligopeptide antibodies in the same buffer but without serum albumin. After washing in the latter buffer, bound antibodies were detected with antibody alkaline phosphatase conjugates (Promega).

Sequence Determination of DNA and Sequence Analysis. The 6:22 cDNA was sequenced using the shotgun sequencing strategy (Messing and Vieira 1982; Deininger 1983). The DNA fragment was self-ligated, sonicated, and cloned into the M13mp9 vector as described (Biggin et al. 1983). The obtained overlapping sequences were aligned and merged with the programs of Staden (1984).

The dideoxy sequencing method was used in combination with ^{35}S-dATP and the modified T7 DNA polymerase (Sequenase, USB). The DNA fragments were separated in buffer gradient gels (Biggin et al. 1983).

Sequence analyses were performed with the programs of Devereaux et al. (1984).

Defined regions of the sp115,140 gene were sequenced using oligodeoxynucleotide primers. These were synthesized on a Pharmacia Gene Assembler.

Oligopeptide Synthesis and Immunological Techniques. The oligopeptide was synthesized in an Applied Biosystems model 430A peptide synthesizer according to a standard program. A phenylacetamidomethyl (PAM) resin and side chain-protected *tert*-butyloxycarbonyl (*t*-Boc) amino acids were used. The peptide was cleaved from the resin and deprotected with hydrogen fluoride. Finally, the crude peptide preparation was purified by HPLC. The identity of the peptide was assessed by amino acid analysis. Purified oligopeptides were coupled to bovine serum albumin with glutaraldehyde, mixed with Freund's adjuvant, and injected

subcutaneously into rabbits. The oligopeptide-specific antibodies were purified by affinity chromatography. The immune sera were passed through two columns of Sepharose 4B to which bovine serum albumin had been coupled. The flow through was then passed through a column containing the oligopeptide coupled to Sepharose 4B. After washing, the bound antibodies were eluted in 4 M $MgCl_2$ and dialyzed against 0.01 M Na_2HPO_4, 0.15 M NaCl, pH 7.4.

Mapping Techniques: Sequencing RNA Templates. The sequence of the 5′ end of the mRNA was determined by direct sequencing of the mRNA according to Geliebter et al. (1986). Specific oligodeoxynucleotides were end-labeled by T4-kinase and ^{32}-P-γ-ATP. Kinased oligonucleotide (0.2–0.4 ng) was hybridized with 15 μg of total gland RNA in 0.4 M NaCl, 0.04 M Pipes (1,4-piperazinediethanesulfonic acid) buffer, pH 6.5, for 2–3 h at the appropriate temperature. The RNA–primer complex was precipitated with ethanol and dissolved in 24 mM Tris-HCl, pH 8.3, 16 mM $MgCl_2$, 8 mM dithiothreitol, 0.4 mM dATP, 0.4 mM dCTP, 0.4 mM dTTP, 0.8 mM dGTP, 100 μg/ml actinomycin D, 20 units of RNasin, and 4 units of reverse transcriptase (Boehringer). The RNA–primer solution was then divided into four tubes and 1 μl of 1 mM ddATP, ddCTP, ddTTP, or ddGTP was added. Incubation was for 45 min at 50°C. Formamide (2 μl) containing 0.3% bromophenol blue and 0.3% xylene cyanol was added, and after incubation at 90°C for 3 min, the samples were electrophoresed in standard sequencing gels.

Mapping Techniques: cDNA Primer Extension. An end-labeled oligodeoxynucleotide primer (21-mer, positions 146–166 in Fig. 3) was hybridized to 15 μg of total gland RNA in 0.4 M NaCl, 0.04 M Pipes buffer, pH 6.5, for 3 h at 50°C. The RNA–primer complex was ethanol precipitated and introduced into the Amersham cDNA kit first strand synthesis reaction. The obtained cDNA was ethanol precipitated, dissolved in formamide containing 0.3% bromophenol blue and 0.3% xylene cyanol, and electrophoresed in a standard sequencing gel. Dideoxy sequencing reactions obtained by oligodeoxynucleotide priming on a fragment subcloned into the Bluescript vector served as a size marker.

Results

Isolation of a Salivary Gland-Specific cDNA

A salivary gland cDNA library was constructed using the λgt10 vector (Huynh et al. 1985). The library was screened with ^{32}P-labeled cDNA reverse transcribed from the same poly(A)$^+$ RNA preparation used for construction of the library, with the aim of identifying cDNA transcripts corresponding to abundant mRNA species. Replica filters were screened in parallel with ^{32}P-labeled probes representing four earlier described genes, BR1 (Wieslander et al. 1982; Case and Byers 1983), BR2.1 (Sümegi et al. 1982), BR2.2 (Case et al. 1983; Wieslander and Lendahl 1983), and BR6 (Lendahl and Wieslander 1984), which are known to produce abundant mRNAs in the gland cells and therefore are highly represented in the cDNA library. Clones hybridizing with the cDNA probe but not with the BR1–BR6 probes were selected.

The tissue-specific presence of the isolated mRNA sequences was then evaluated. The salivary glands were mechanically removed from larvae and poly(A)$^+$ RNA prepared from all remaining tissues. ^{32}P-labeled single-stranded cDNA was reverse transcribed and used to probe the isolated cDNA clones (not shown). All were negative in agreement with the presence of the corresponding mRNAs only in the salivary gland cells. As a control, a cloned globin gene (Antoine and Niessing 1984) was readily detected using the same cDNA probe.

Several of the cDNAs were subcloned into a plasmid vector and ^{3}H-labeled RNA transcripts were prepared and hybridized to salivary gland chromosomes in situ. One cDNA, called 6:22, hybridized to region 17 on chromosome I (Fig. 1), a locus that does not form a Balbiani ring.

The size of the corresponding mRNA was determined by hybridizing the 6:22 cDNA to Northern blots of total salivary gland RNA (Fig. 2). The relative mobility of the mRNA corresponds to a size of 3.7 kb.

We conclude that the 6:22 cDNA represents an as yet unidentified gene, which is transcribed into an abundant mRNA only in the salivary gland cells and therefore presumably encodes one of the secretory proteins.

The 6:22 cDNA Sequence and the Identification of the Encoded Protein

The 3.1-kb-long cDNA was cleaved out from the λgt10 vector, and without any intermediate subcloning, its sequence was determined using the shotgun sequencing strategy. Upon aligning the obtained random sequences, it became clear that a substantial part of the cDNA consisted of 42-bp-long repeats organized in one long central array. These repeats were similar enough to prevent proper alignment, and therefore it was possible to put together only overlapping sequences covering the two ends of the cDNA into continuous sequences (Fig. 3). At both the 5′ and the 3′ ends a few repeats could be included because they were parts of sequences that could be read from one single DNA fragment, starting in the nonrepetitive ends and extending a number of repeats toward the center of the gene. The central gap between the 5′ and the 3′ end of the cDNA consists only of an array of 42-bp repeats. No other type of sequence was present among the random fragments sequenced, together comprising five times the length of the cDNA insert.

The 6:22 cDNA lacked about 0.6 kb of the mRNA sequence. At the 5′ end 121 of the remaining 125 bp at the 5′ end of the mRNA were determined by using the mRNA as template and two oligodeoxynucleotides as primers in sequencing reactions (Fig. 4). At the 3′ end, the cDNA started in the 3′ untranslated region 120 bp 3′ of the stop codon. Additional

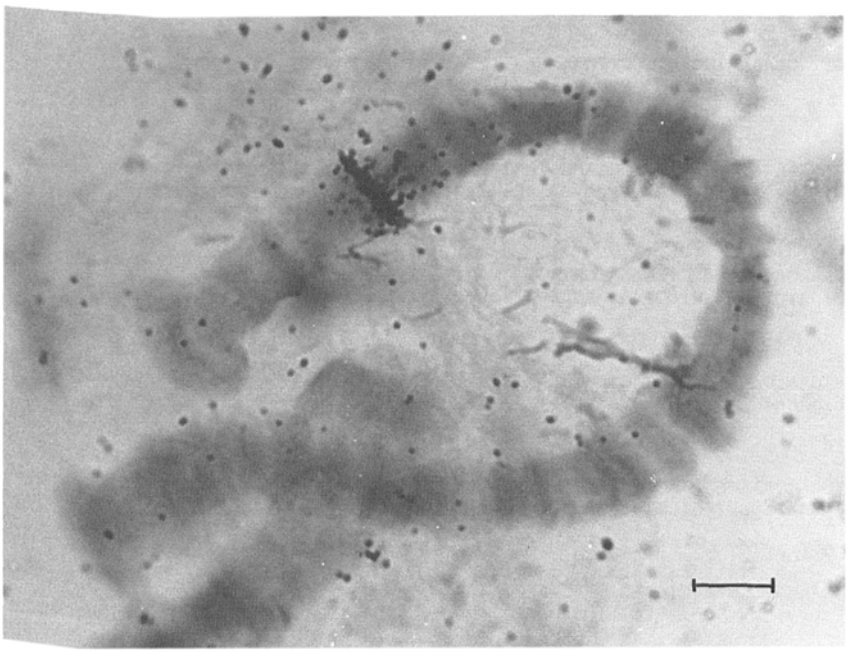

Fig. 1. In situ hybridization of ^{3}H-labeled RNA transcripts from the 6:22 cDNA to salivary gland polytene chromosomes. Specific hybridization to locus 17 on chromosome I is shown. The bar represents 10 μm.

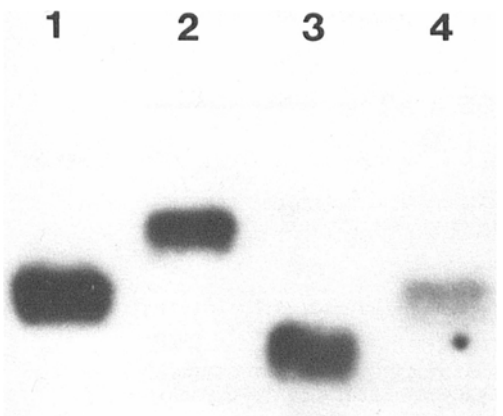

Fig. 2. Hybridization of the 6:22 cDNA to a Northern blot of salivary gland RNA. Total gland RNA and linearized plasmid DNA molecules of known lengths were denatured in 2.2 M formaldehyde, 50% formamide; electrophoresed in 0.8% agarose containing formaldehyde; and transferred to a nylon membrane. The membrane was divided and incubated with ^{32}P-labeled plasmid DNA (lanes 1–3) or ^{32}P-labeled 6:22 cDNA (lane 4). The linear plasmid DNA size markers are 3.8 kb (lane 1), 5.4 kb (lane 2), and 3 kb (lane 3).

sequences extending in the 3′ direction were obtained from a genomic fragment. Two overlapping poly(A) addition sequences are present around 270 bp from the stop codon. Because these are located at approximately the expected site and no other such sequences are present within reasonable distance, we assume that one or both of these poly(A) addition sequences are used.

The coding strand was determined by hybridizing oligodeoxynucleotides, representing the two DNA strands, to Northern blots of total salivary gland RNA (not shown). In the coding strand, only one reading frame is open in all the repeats, and in Fig. 3 the corresponding amino acid sequence is given.

In order to demonstrate if any of the salivary gland secretory proteins is encoded by the cDNA sequence, we made an oligopeptide corresponding to parts of the conceptually translated repeat sequence. Antibodies directed toward this oligopeptide, Lys–Pro–Gly–Gly–Lys–Gly–Pro–Asn–Lys–Gly–Lys–Ser, were raised in immunized rabbits and purified by affinity chromatography. In Western blots, the antibodies detected two salivary gland secretory proteins with apparent relative molecular masses of 115,000 and 140,000 (Fig. 5).

The binding to both proteins is specific and can be competed out by the oligopeptide. From the cDNA sequence complemented at the 5′ end, the calculated molecular weight of the complete protein is approximately 105,000. The slower relative mobility recorded may be due either to posttranslational modifications or to the unusual amino acid composition. There are several possibilities to explain why we detect two bands. First, the bands may represent two differently modified versions of the same protein. Second, they may reflect the presence of two alternative transcription start sites and/or alternatively spliced mRNAs. This alternative appears unlikely because we detect only one type of transcript and one transcription start site (see below). Third, the two bands may represent two proteins encoded in separate genes but sharing the same epitope. The latter alternative is supported by a recent finding where antibodies raised against an oligopeptide derived from a partial cDNA clone representing a closely related but apparently not

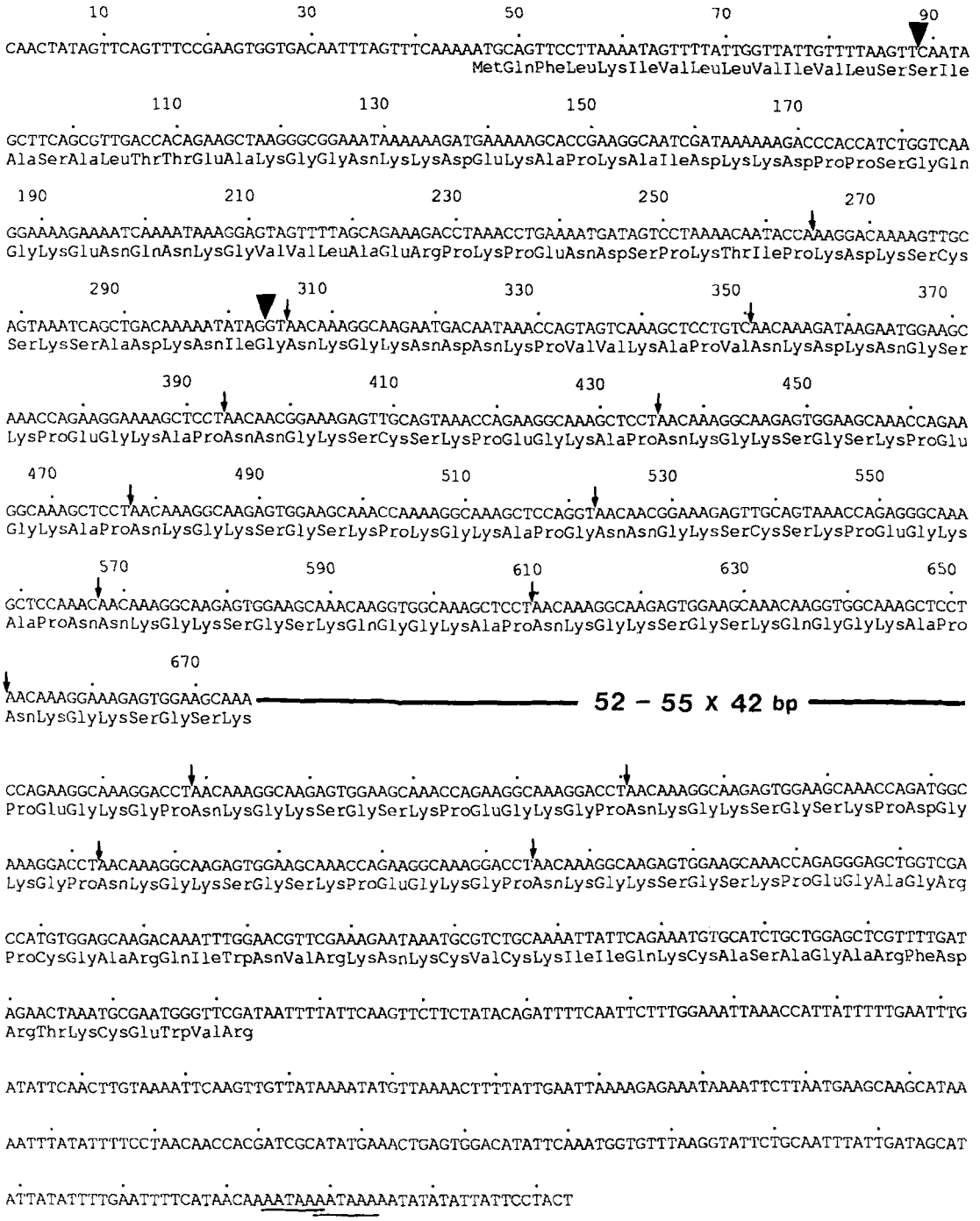

Fig. 3. Nucleotide and corresponding amino acid sequence of the sp115,140 gene mRNA. The sequence of the 5′ and 3′ ends of the mRNA is shown. The gap consists of an array of the 42-bp repeat. The positions of the two introns in the gene are indicated by thick arrows. Thin arrows mark the 42-bp repeats. Two possible poly(A) addition sequences are underlined.

identical gene identified only the 140-kd protein (Dignam et al. 1989).

The primary structure of the protein, deduced from the cDNA sequence is simple in design. In the large central core block, the 14-amino acid repeat is dominated by lysine and glycine residues (8/14). This part of the protein is quite hydrophilic, and the lysine residues are regularly spaced, (KxKxxx-KxxxKxxx) times *n*. At the C-terminal, there is a 31-residue-long nonrepetitive sequence containing four cysteine residues. At the N-terminal, the core block is preceded by a region with 60 residues, which in composition is quite similar to the core block, but no obvious repeats are present. This part does not contain cysteine residues, but in the very beginning of the core block, repeat number one, four, and

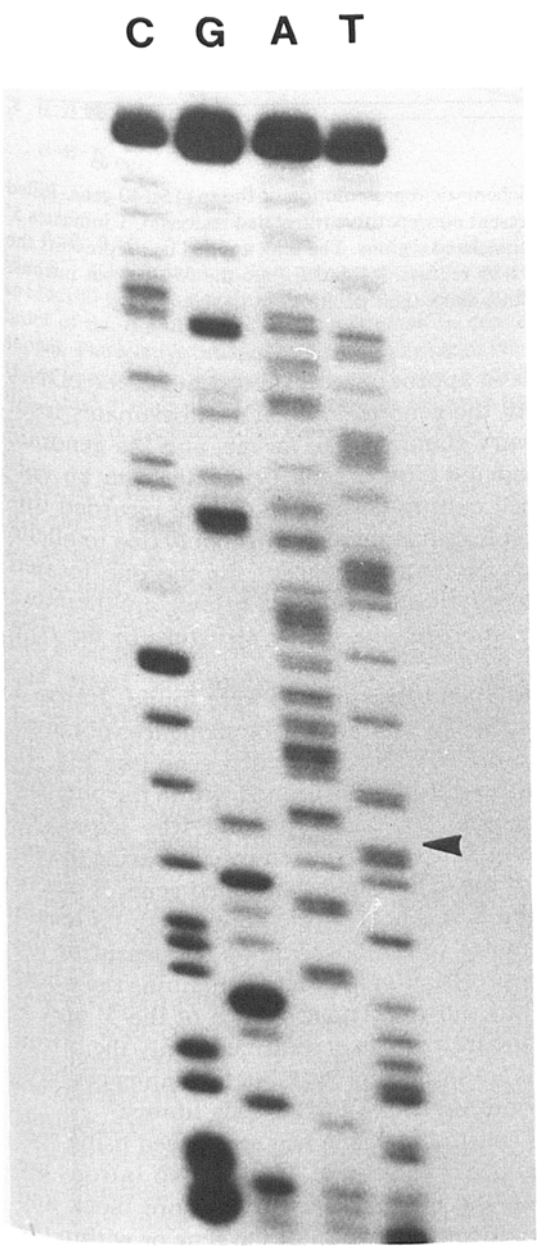

Fig. 4. Autoradiogram of sequence reactions using the sp115,140 gene mRNA as template. The sequence from a defined oligonucleotide primer to the 5′ end of the mRNA is shown. The arrow marks the position of intron 1.

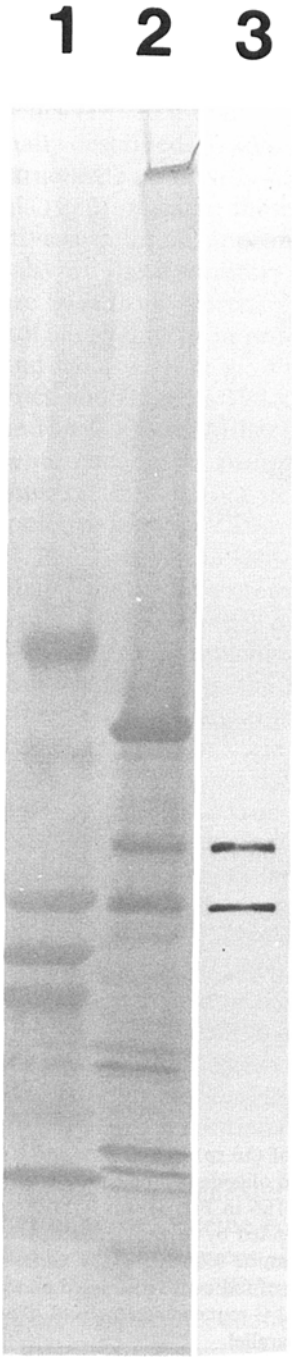

Fig. 5. Western blot of salivary gland secretory proteins. Gland lumen proteins were separated in a 3–20% concave exponential polyacrylamide gel and transferred to a nitrocellulose filter. Antibodies directed against an oligopeptide specified by the 6:22 cDNA sequence specifically bind to two protein bands with apparent relative molecular masses of 115,000 and 140,000. Lane 1, size markers with molecular weights of from top to bottom 205, 116, 97.4, 66, and 45 kd, stained by amido black. Lane 2, salivary gland lumen proteins stained by amido black. Lane 3, identical to lane 2, incubated with the antibodies and stained with antibody alkaline phosphatase conjugates.

seven each have one cysteine residue. Overall the protein thus has a large central repetitive and hydrophilic region without any cysteine residues. This domain is surrounded by short nonrepetitive regions and three to four closely spaced cysteine residues.

The Structure of the sp115,140 Gene

The gene corresponding to the 6:22 cDNA was isolated from an EMBL 4 *C. tentans* genomic library. Until data are available to settle if the 115-kd or 140-kd protein (or both) is encoded by this gene, we will refer to it as the sp115,140 gene. The entire gene was present in a 7.2-kb EcoRI fragment, and the gene structure was established by a combination of Southern blot analysis and determination of the

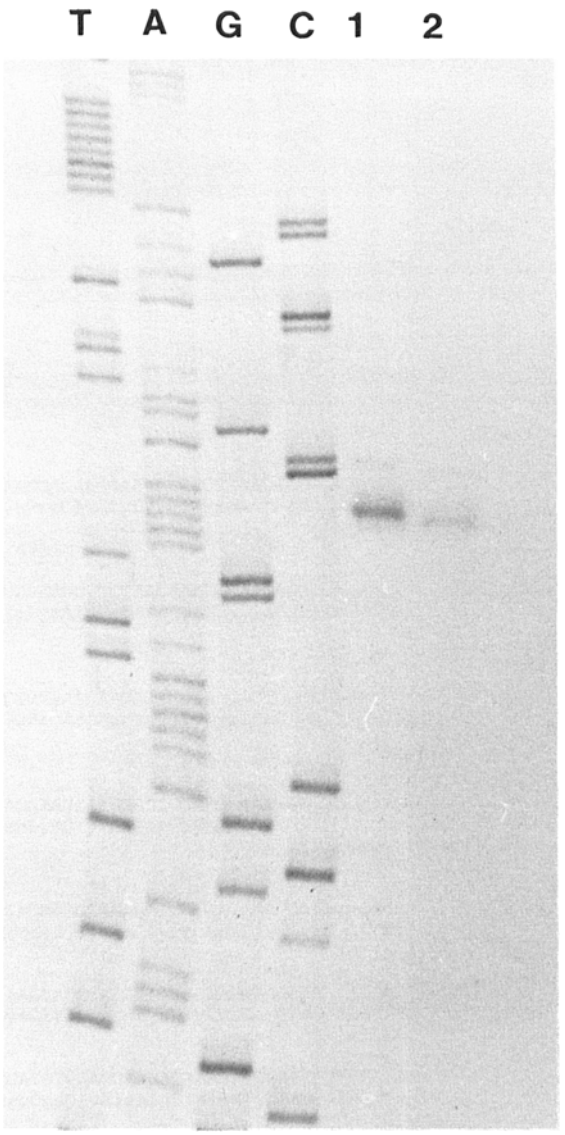

Fig. 6. Mapping of the 5′ end of the sp115,140 gene mRNA by cDNA extension. A ^{32}P-labeled oligodeoxynucleotide primer (corresponding to positions 146–166 in Fig. 3) was hybridized to total salivary gland RNA, extended by reverse transcriptase, and loaded onto a 6% polyacrylamide sequencing gel. Lanes 1 and 2 represent different amounts of radioactivity loaded on the gel. As size markers, T, A, G, and C sequencing reactions of an unrelated sequence were run in parallel.

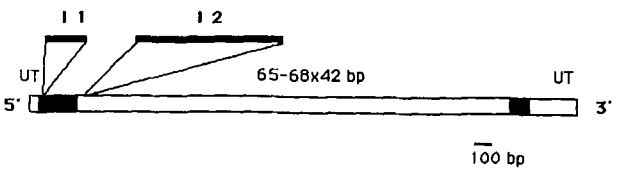

Fig. 7. Schematic representation of the sp115,140 gene. Filled boxes represent nonrepetitive translated regions. UT indicates 5′ and 3′ untranslated regions. The long unfilled box represents the array of 42-bp repeats. I 1 and I 2 are the two cut out introns. The thin lines show their positions in the gene.

sequence at defined regions of the gene. The start point of transcription was determined by cDNA extension priming (Fig. 6) in combination with sequencing a genomic fragment covering the transcription start point. In Fig. 7 the gene is depicted schematically. The size and noninterrupted repetitive structure of the central core block were shown by measuring and comparing the length of the core block in the 6:22 cDNA and in the genome as well as in the cloned gene. The central core block was shown to be 2.8 kb long in the 6:22 cDNA and about 120 bp shorter in both the genome and the cloned gene (not shown). The length of the core block corresponds to approximately 68 repeats in the cDNA and 65 in the genome. The cDNA originates from the salivary gland cells of larvae, and the genomic DNA and the cloned gene originate from an epithelial cell culture (Wyss 1982). The recorded difference in length is therefore probably due to allelic variation. No introns, except for the one located within repeat number one (see below), are therefore likely to interrupt the repeat structure in the core block.

At the 5′ end two introns were found. Intron 1 was identified by comparing the sequences obtained from the mRNA and the cloned gene (see Fig. 4). This 268-bp intron is positioned within the putative signal peptide coding region and was sequenced completely. Intron 2, detected by comparing the sequence of the cDNA and the cloned gene, is placed within the first repeat of the core block. Its length was estimated to be about 950 bp by measuring the length of a restriction fragment containing the whole intron (not shown). The sequence of the 5′ and 3′ ends of the intron was determined. Only the intron positions are shown in Fig. 3, but the sequences have been submitted to the EMBL data library.

The 3′ part of the gene was sequenced using specific oligodeoxynucleotide primers. No introns are present at the border between the core block and the 3′ nonrepetitive region of the gene or within the latter region.

Sequence Homogeneity within the sp115,140 Gene

The repeats in the sp115,140 gene are highly similar to each other, but two types of sequence divergence are present. First, the central repeats come in two main versions, differing by a few base pair substitutions. These are present in only 4 of the 14 codons of the repeat, and in two cases they lead to amino acid substitutions (Ala–Gly and Gly–Glu) (see Fig. 3). The base pair substitutions occur in essentially only two combinations and the repeats are therefore of two kinds, called A and B. The distribution of these along the gene could be inferred from random fragments containing several (4–7) repeats whose complete sequences were determined in a single sequence reaction. Judging from these (Fig. 8), the A

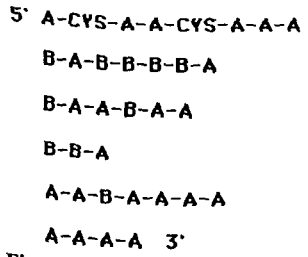

Fig. 8. Distribution of the two repeat unit variants in the sp115,140 gene. Six regions from the core block in which the order of the A and B type repeat unit could be determined are shown. The relative order of the four middle regions in the figure is not known. The two regions marked 5' and 3' are located at the 5' and 3' ends of the core block. Cys indicates repeat units at the 5' end, containing one cysteine codon each.

version is the most common one, and the B version appears as single repeats within a cluster of A repeats or as small B clusters.

Second, variant repeats are present at both the 5' and 3' ends. At the 3' end, the next to last repeat has base pair substitutions unique to itself, and the last repeat can only be identified as a repeat in its 5' half. At the 5' end, several repeats are different. They have a number of base pair substitutions, in three repeats introducing one cysteine codon. Four repeats have one extra codon, and in the first repeat, intron 1 is located right within the extra codon of that repeat.

In addition, at the 5' end the repetitive part does not start abruptly at a defined point. It is in fact hard to decide the exact border between the 5' nonrepetitive sequences and the first repeat. The overall amino acid composition of the ~60-amino acid-long nonrepetitive part is almost identical to that of the repeats. In addition, two regions with sequences similar to each other are discernible within the nonrepetitive region, which are also similar in part to the sequence of the repeats. One of these regions codes for three lysines that are spaced three amino acids apart, exactly as in the repeats. These similarities may be a sign of a sequence relationship between the repeat sequences and the 5' nonrepeated sequences, extending all the way to the 5' end of the coding part of the gene.

Discussion

We have described the gene for 1 of the approximately 15 different secretory proteins synthesized by the salivary gland cells in *C. tentans*. The most prominent feature of this sp115,140 gene is its repetitive structure. A central block with 65–68 copies of a 42-bp repeat arranged in tandem is surrounded by short nonrepetitive 5' and 3'-translated sequences.

Five of the other secretory protein-encoding genes in *C. tentans*, the BR1 (Wieslander et al. 1982; Case and Byers 1983), BR2.1 (Sümegi et al. 1982), BR2.2 (Case et al. 1983; Wieslander and Lendahl 1983), BR6 (Lendahl and Wieslander 1984), and sp195 genes (Dreesen et al. 1985), have to date been partially described. In addition, the complete BR3 gene structure has recently been determined (Paulsson et al. 1989). Because these genes also exhibit a repetitive structure, it appears that all the genes encoding salivary gland secretory proteins larger than 115 kd are built from repetitive sequence elements. This is not surprising, as in prokaryotes (Uhlén et al. 1984) and eukaryotes (e.g., Yamada et al. 1980; Muskavitch and Hogness 1982; Kataoka et al. 1985; Eckert and Green 1986; Miller et al. 1986) in general, many genes encoding a protein whose function involves multiple interactions of the same nature along the molecule are built from repetitive sequences.

The *C. tentans* salivary glands excrete protein fibers. All of the secretory proteins presumably interact with each other during the assembly of these fibers and/or in attaching the fibers to each other as they are spun into the fiber network of the larval tube. In the corresponding genes the repetitive structures appear to be variations on a common theme. In the closely related BR1, BR2.1, BR2.2, and BR6 genes encoding the large sp-I a–d proteins (Kao and Case 1985; Botella et al. 1988), two ~100-bp-long regions, the C-region and the SR-region, are repeated in an alternating fashion throughout at least 20–25 kb of the 35-kb-long genes (see Fig. 9) (Pustell et al. 1984; Wieslander et al. 1984; Grond et al. 1987). The C-region contains four conserved cysteine codons. The SR-regions all maintain the theme of a proline residue preceded by a positively charged residue (a lysine or arginine) and followed by a negatively charged residue (a phosphorylated serine or a glutamic acid).

In the sp195 gene the 75-bp-long repeat contains two or three cysteine codons as well as short segments with sequences similar to the SR-regions of the large BR genes (Dreesen et al. 1985 and unpublished results), i.e., elements of both the C- and SR-regions.

In the BR3 gene (Paulsson et al. 1989), a repeat structure is present with conserved and regularly spaced cysteine codons, corresponding to only the C-region.

In contrast, we demonstrate that the sp115,140 gene has a simple repeat structure essentially devoid of cysteine codons. In fact, the amino acid sequence of the 42-bp repeat is in large part identical to stretches of the subrepeats in the SR-regions of the large BR genes (see Fig. 9). The sp115,140 gene therefore looks like the core block of a large BR gene from which the C-regions have been removed, leaving many juxtaposed SR-regions.

We conclude that the protein components of the

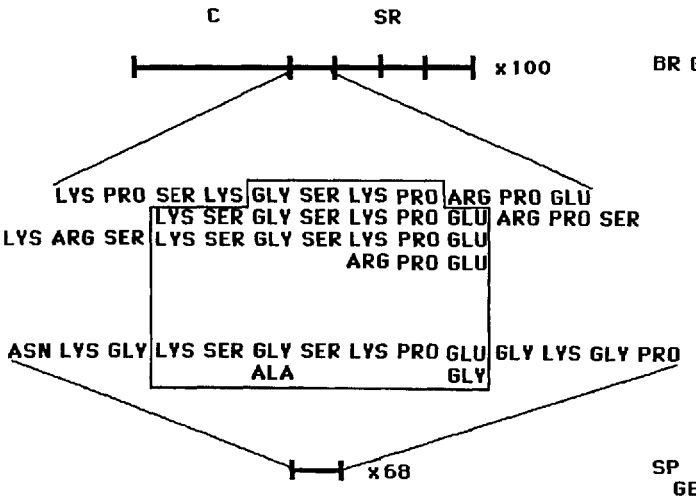

Fig. 9. Relationship between the 42-bp repeat unit of the sp115,140 gene and the repeat unit of the large BR genes. One repeat unit from the large BR genes (top) and one from the sp115,140 gene (bottom) are shown schematically. The amino acid sequence of the sp115,140 gene 42-bp repeat unit and one subrepeat from within the SR region of the repeat units in three different large BR genes are shown. The BR1λ and BR1β repeat units are located within the same gene (unpublished results). Boxed sequences show regions of identity.

gland secretion so far described all have repetitive structures and that the repeats are similar between the proteins. Two repetitive components, one containing cysteine and one containing charged proline, appear to be the functional units, and one protein may have both of these units or, as shown by the sp115,140 and the BR3 gene, either one of them.

Repeat structures with similar sequence motifs in various genes could have evolved from separate origins by convergent evolution. The alternative and more likely explanation is that the genes are related to each other by common ancestry. A common origin of the sp115,140 gene and the earlier described secretory protein genes would then require that the repeat structures in these genes arose by reduplications of sequences originating from a common ancestor and/or appeared as a result of remodelings within an already existing repeat structure. Continuous selection for repeat structures that have allowed proper interactions between the corresponding proteins must then have been operating, and the sp115,140 gene could represent one alternative outcome of such events.

The structure of the sp115,140 gene exhibits several features which suggest that remodelings of the repeat structure necessary for the latter evolutionary model do occur. In the core block all repeats are almost identical. They therefore have to be homogenized efficiently and evolve concertedly, similar to other tandem repeats (Arnheim 1983). The three types of sequence divergence that are present in the gene suggest that the repeat structure does change. First, the pattern of the A and B type of repeats in the core block is in agreement with the idea that repeat variants may be introduced and spread within an existing array of repeats.

Second, the presence of variant repeats at the ends of the core block, in particular at the 5′ end, shows that these repeats are not efficiently homogenized together with the central repeats. This may occur because homogenization involves recombination between unequally aligned homologous DNA molecules (Barker et al. 1988; Höög et al. 1988) and indicates that the repeat structure has changed during evolution.

Third, the similarities between the nonrepetitive 5′ end and the core block repeats may be a sign of a sequence relationship between the repeat sequences and the 5′ nonrepeated sequences extending all the way to the 5′ end of the coding part of the gene.

The repetitive structure of the sp115,140 gene therefore has most likely been remodeled during evolution. This is in agreement with the observation that repeat arrays in other coding genes change in sequence and number of repeats over short evolutionary time scales, as, e.g., in the involucrine gene of primates (Tseng and Green 1988), the large BR genes (Lendahl et al. 1987) and genes for high molecular weight glutenin subunits (Goldsbrough 1988), the silk fibroin gene (Manning and Gage 1980), the glue protein genes in *Drosophila* (Muskavitch and Hogness 1982), S-antigen genes in plasmodia (Cowman et al. 1985), and the apolipoprotein(a) gene (McLean et al. 1987). Repeat arrays therefore are subject to frequent changes, and these have been attributed to nonreciprocal recombinations after unequal alignment of the repeat structures (Smith 1976; Maeda and Smithies 1986), slipped-mispairing during DNA replication (Levinson and Gutman 1987), and gene conversion (Jackson and Fink 1981; Klein and Petes 1981).

We conclude that different repeat structures may have evolved from a common origin and resulted in the sp115,140 gene and the remaining set of functionally coupled secretory protein genes. This conclusion receives support from observations that link the SR-type and C-type regions to each other and make it less likely that they arose independently from unrelated sequences several times during evo-

lution. The sp115,140 gene has only the SR-type region and the BR3 gene only the C-type region. At the same time, the four large BR genes have the SR- and C-type regions and both appear to have evolved from within the same repeat structure in an ancestor gene (Pustell et al. 1984; Höög et al. 1988). In addition, the direct interaction between the various secretory proteins can be understood if their genes arose by gene duplications and subsequent divergence. It is more difficult to see how such a functional connection was established between originally completely unrelated genes. The former evolutionary history is common. The globin gene family (Weatherall and Clegg 1979), the acetylcholine receptor (Mishina et al. 1985), and the immunoglobulin supergene family (Hood et al. 1985) are such examples. By analogy, it is possible that the evolution of the different secretory protein genes has resulted in a better larval tube, which perhaps also can attain slightly different properties depending on the combination of genes used at a given time.

Acknowledgments. We thank Kerstin Bernholm for technical assistance. This work was supported by the Swedish Natural Science Foundation, the Swedish Medical Research Council, Magnus Bergvalls Stiftelse, and Karolinska Institutet. The peptide synthesis was supported by the Swedish Cancer Society (project no. 1806), the Swedish Medical Research Council (project no. 1010), and Stiftelsen Bengt Lundqvists Minne.

References

Antoine M, Niessing J (1984) Intron-less globin genes in the insect *Chironomus thummi thummi.* Nature 310:795–798

Arnheim N (1983) Concerted evolution of multigene families. In: Nei M, Koehn RK (eds) Evolution of genes and proteins. Sinauer, Sunderland MA, pp 38–61

Barker RF, Harbered NP, Jarvis MG, Flavell RB (1988) Structure and evolution of the intergenic region in a ribosomal DNA repeat unit of wheat. J Mol Biol 210:1–17

Benton WD, Davis RW (1977) Screening λgt recombinant clones by hybridization to single plaques in situ. Science 196:180–182

Biggin MD, Gibson TJ, Hong GF (1983) Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc Natl Acad Sci USA 80:3963–3965

Botella L, Grond C, Saiga H, Edström J-E (1988) Nuclear localization of a DNA-binding C-terminal domain from Balbiani ring coded secretory protein. EMBO J 7:3881–3888

Burnette WN (1981) Western blotting: electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. Anal Biochem 112: 195–203

Case ST (1986) Correlated changes in steady-state levels of Balbiani ring mRNAs and secretory polypeptides in salivary glands of *Chironomus tentans.* Chromosoma 94:483–491

Case ST, Byers MR (1983) Repeated nucleotide sequence arrays in Balbiani ring 1 of *Chironomus tentans* contain internally nonrepeating and subrepeating elements. J Biol Chem 258: 7793–7799

Case ST, Summers RL, Jones AG (1983) A variant tandemly repeated nucleotide sequence in Balbiani ring 2 of *Chironomus tentans.* Cell 36:555–562

Cowman AF, Saint RB, Coppel RL, Brown GV, Anders RF, Kemp DJ (1985) Conserved sequences flank variable tandem repeats in two S-antigen genes of *Plasmodium falciparum.* Cell 40:775–783

Deininger P (1983) Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. Anal Biochem 129:216–223

Devereaux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 12:387–395

Dignam SS, Yang L, Lezzi M, Case ST (1989) Identification of a developmentally regulated gene for a 140-kDa secretory protein in salivary glands of *Chironomus tentans* larvae. J Biol Chem 264:9444–9452

Dreesen TD, Bower JR, Case ST (1985) A second gene in a Balbiani ring: *Chironomus* salivary glands contain a 6.6 kb poly(A)$^+$ RNA that is transcribed from a hierarchy of tandem repeated sequences in Balbiani ring 1. J Biol Chem 260:11824–11830

Dretzen G, Bellard M, Sassone-Corsi P, Chambon P (1981) A reliable method for the recovery of DNA fragments from agarose and acrylamide gels. Anal Biochem 112:295–298

Eckert RL, Green H (1986) Structure and evolution of the human involucrin gene. Cell 46:583–589

Edström J-E, Rydlander L, Francke C (1980) Concomitant induction of a Balbiani ring and a giant secretory protein in *Chironomus* salivary gland. Chromosoma 81:115–124

Feinberg AP, Vogelstein B (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem 132:6–13

Geliebter J, Zeff RA, Melvold RW, Nathenson SG (1986) Mitotic recombination in germ cells generated two major histocompatibility complex mutant genes shown to be identical by RNA sequence analysis: K^{bm9} and K^{bm6}. Proc Natl Acad Sci USA 83:3371–3375

Goldsbrough AP, Robert L, Schnick D, Flavell RB (1988) Molecular comparisons between bread-making-quality determining high molecular weight glutenin subunits of wheat gluten—evidence for the nature of good and poor quality at the protein level. In: Miller TE, Coebner RMD (eds) Proceedings of the VII International Wheat Genetics Symposium. Institute for Plant Science Research, Cambridge, pp 727–733

Grond C, Saiga H, Edström J-E (1987) The sp-I genes in the Balbiani rings of *Chironomus* salivary glands. In: Hennig W (ed) Results and problems in cell differentiation, vol 14. Springer-Verlag, Berlin, pp 69–80

Grossbach U (1977) The salivary gland of *Chironomus* (Diptera): a model system for the study of cell differentiation. In: Beermann W (ed) Results and problems in cell differentiation, vol 8. Springer-Verlag, Berlin, pp 147–196

Gross-Bellard M, Oudet P, Chambon P (1973) Isolation of high-molecular-weight DNA from mammalian cells. Eur J Biochem 36:32–38

Gubler U, Hoffman BJ (1983) A simple and very efficient method for generating cDNA libraries. Gene 25:263–269

Hood L, Kronenberg M, Hunkapiller T (1985) T cell antigen receptors and the immunoglobulin supergene family. Cell 40: 225–229

Höög C, Wieslander L, Daneholt B (1988) Terminal repeats in long repeat arrays are likely to reflect the early evolution of Balbiani ring genes. J Mol Biol 200:655–664

Huynh TV, Young RA, Davis RW (1985) Constructing and screening cDNA libraries in λgt10 and λgt11. In: Glover DM (ed) DNA cloning—a practical approach, vol I. IRL Press, Oxford, pp 49–78

Jackson FR, Fink GR (1981) Gene conversion between duplicated genetic elements in yeast. Nature 292:306–311

Kao, W-Y, Case ST (1985) Individual variations in the content of giant secretory polypeptides in salivary glands of *Chironomus*. J Cell Biol 101:1044–1051

Kataoka T, Broek D, Wigler M (1985) DNA sequence and characterization of the *S. cerevisiae* gene encoding adenylate cyclase. Cell 43:493–505

Klein HL, Petes TD (1981) Intrachromosomal gene conversion in yeast. Nature 289:144–148

Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of the bacteriophage T4. Nature 227: 680–685

Lehrach H, Diamond D, Wozney JM, Boedker H (1977) RNA molecular weight determinations by gel electrophoresis under denaturing conditions, a critical reexamination. Biochemistry 16:4743–4751

Lendahl U, Wieslander L (1984) Balbiani ring 6 gene in *Chironomus tentans*: a diverged member of the Balbiani ring gene family. Cell 36:1027–1034

Lendahl U, Wieslander L (1987) Balbiani (BR) genes exhibit different patterns of expression during development. Dev Biol 121:130–138

Lendahl U, Saiga H, Höög C, Edström J-E, Wieslander L (1987) Rapid and concerted evolution of repeat units in a Balbiani ring gene. Genetics 117:43–49

Levinson G, Gutman GA (1987) High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. Nucleic Acids Res 15:5323–5338

Maeda N, Smithies OA (1986) The evolution of multigene families: human haptoglobin genes. Annu Rev Genet 20:81–108

Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor NY

Manning FR, Gage PL (1980) Internal structure of the silk fibroin gene of *Bombyx mori*. II. Remarkable polymorphism of the organization of crystalline and amorphous coding sequences. J Biol Chem 255:9451–9457

McLean JW, Tomlinson JE, Kuang W-J, Eaton DL, Chen EY, Fless GM, Scanu AM, Lawn RM (1987) cDNA sequences of human apolipoprotein(a) is homologous to plasminogen. Nature 300:132–137

Messing J, Vieira J (1982) A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments. Gene 19:269–276

Miller HL, Howard RJ, Carter R, Good MF, Nussenzweig V, Nussenzweig RS (1986) Research towards malaria vaccines. Science 234:1349–1356

Mishina M, Takahashi T, Takai T, Kurasaki M, Fukuda K, Numa S (1985) Role of acetylcholine receptor subunits in gating of the channel. Nature 318:538–543

Muskavitch MAT, Hogness DS (1982) An expandable gene that encodes a *Drosophila* glue protein is not expressed in variants lacking remote upstream sequences. Cell 29:1041–1051

Paulsson G, Lendahl U, Galli J, Ericsson C, Wieslander L (1990) The Balbiani ring 3 gene in *Chironomus tentans* has a diverged repetitive structure split by many introns. J Mol Biol 277: 337–349

Pustell J, Kafatos F, Wobus U, Bäumlein H (1984) Balbiani ring DNA: sequence comparisons and evolutionary history of a family of hierarchically repetitive protein coding genes. J Mol Evol 20:281–295

Rydlander L, Edström J-E (1980) Large size nascent protein as dominating component during protein synthesis in *Chironomus* salivary glands. Chromosoma 81:101–113

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191:528–535

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503–517

Staden R (1984) A computer program to enter DNA gel reading data into a computer. Nucleic Acids Res 12:499–503

Sümegi J, Wieslander L, Daneholt B (1982) A hierarchic arrangement of the repetitive sequences in the Balbiani ring 2 gene of *Chironomus tentans*. Cell 30:579–587

Thomas PS (1980) Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. Proc Natl Acad Sci USA 77:5201–5205

Tseng H, Green H (1988) Remodelling of the involucrine gene during primate evolution. Cell 54:491–496

Uhlén M, Guss B, Nilsson B, Gatenbeck L, Philipson L, Lindberg M (1984) Complete sequence of the staphylococcal gene encoding protein A. J Biol Chem 259:1695–1702

Wallace RB, Miyada CG (1987) Oligonucleotide probes for the screening of recombinant DNA libraries. In: Berger SL, Kimmel AR (eds) Methods in enzymology, vol 152. Academic Press, San Diego, pp 432–442

Weatherall DJ, Clegg JB (1979) Recent developments in the molecular genetics of human hemoglobin. Cell 16:467–479

Wieslander L, Lendahl U (1983) The Balbiani ring 2 gene in *Chironomus tentans* is built from two types of tandemly arranged major repeat units with a common evolutionary origin. EMBO J 2:1169–1175

Wieslander L, Sümegi J, Daneholt B (1982) Evidence for a common ancestor sequence for the Balbiani ring 1 and Balbiani ring 2 genes in *Chironomus tentans*. Proc Natl Acad Sci USA 79:6956–6960

Wieslander L, Höög C, Höög J-O, Jörnvall H, Lendahl U, Daneholt B (1984) Conserved and nonconserved structures in the secretory proteins encoded in the Balbiani ring genes of *Chironomus tentans*. J Mol Evol 20:304–312

Wyss C (1982) *Chironomus tentans* epithelial cell lines sensitive to ecdysteroids, juvenile hormone, insulin and heat shock. Exp Cell Res 139:297–307

Yamada Y, Avvedimento VE, Mudryj M, Ohkubo H, Vogeli G, Irani M, Pastan I, de Crombrugghe B (1980) The collagen gene: evidence for its evolutionary assembly by amplification of a DNA segment containing an exon of 54 bp. Cell 22:887–892

Received August 28, 1989/Revised and accepted January 6, 1990