Molecular Biology of the Cell
Vol. 6, 649–659, June 1995

A Large Multigene Family Codes for the Polypeptides of the Crystalline Trichocyst Matrix in *Paramecium*

Luisa Madeddu, Marie-Christine Gautier, Laurence Vayssié, Abdellah Houari, and Linda Sperling*

†Centre de Génétique Moléculaire, Centre National de la Recherche Scientifique, 91198 Gif-sur-Yvette Cedex, France

Submitted January 11, 1995; Accepted March 30, 1995
Monitoring Editor: Joseph Gall

The secretory granules (trichocysts) of *Paramecium* are characterized by a highly constrained shape that reflects the crystalline organization of their protein contents. Yet the crystalline trichocyst content is composed not of a single protein but of a family of related polypeptides that derive from a family of precursors by protein processing. In this paper we show that a multigene family, of unusually large size for a unicellular organism, codes for these proteins. The family is organized in subfamilies; each subfamily codes for proteins with different primary structures, but within the subfamilies several genes code for nearly identical proteins. For one subfamily, we have obtained direct evidence that the different members are coexpressed. The three subfamilies we have characterized are located on different macronuclear chromosomes. Typical 23–29 nucleotide *Paramecium* introns are found in one of the regions studied and the intron sequences are more variable than the surrounding coding sequences, providing gene-specific markers. We suggest that this multigene family may have evolved to assure a microheterogeneity of structural proteins necessary for morphogenesis of a complex secretory granule core with a constrained shape and dynamic properties: genetic analysis has shown that correct assembly of the crystalline core is necessary for trichocyst function.

INTRODUCTION

The generation of multigene families is a corollary to the increase in complexity of genomes and organisms in the course of evolution (for a review, see Ohta, 1991). Gene duplications may become fixed in a population as the duplicated genes acquire beneficial new functions or patterns of expression. Alternatively, multiple copies of functionally equivalent genes have been suggested to provide a buffer against deleterious mutations (Clark, 1994).

Very large multigene families (tens or hundreds of members) have evolved in a number of instances, most often providing a finely differentiated repertoire of substrate specificities or of "sensors" in response to a complex environment. Some examples are the genes coding for Igs (Tonegawa, 1983), odor receptors (Buck and Axel, 1991), and cytochrome P450s (Black and Coon, 1987). In most of these cases, a hierarchy of control mechanisms assures the expression of a single member of the family in a given cell at a given time. Other examples of large multigene families are insect chorion genes (Goldsmith and Kafatos, 1984) and plant storage protein genes (Wilson and Larkins, 1984). In both of these cases there is in part, diversity of expression of the genes at different stages of development of egg shell or seed kernel, some diversity of the encoded structural proteins, and some redundancy to assure synthesis of large amounts of the proteins. Finally, there is the extreme case of histone and rRNA genes, characterized by a nearly total lack of diversity; here high copy number has evolved to provide the cell with large amounts of identical gene products.

In the course of our studies of trichocyst secretion we have gathered evidence for the existence of a large multigene family in *Paramecium* (reported in preliminary form in Madeddu *et al.*, 1994). *Paramecium* and

* Corresponding author.
† Associated with the Université Pierre et Marie Curie.

© 1995 by The American Society for Cell Biology

other ciliates are among the few unicellular organisms to have regulated secretion, a function generally restricted to specialized cells of metazoans. Each cell bears some 1000 architecturally complex secretory granules, known as trichocysts, docked at specialized cortical sites ready to rapidly and synchronously release their contents in response to extracellular stimuli (see Adoutte, 1988 for a review). The trichocysts have a highly constrained shape reflecting the crystalline organization of the granule protein contents, which undergoes an irreversible structural transition at exocytosis, upon contact with the external medium (Sperling *et al.*, 1987).

The data presented here provide direct evidence that the proteins of the crystalline trichocyst core are coded by a family of some 100 coexpressed genes. Not only is this family of unusually large size for a unicellular organism, but it fails to enter any of the aforementioned configurations of a large multigene family. We suggest that this family may have evolved to generate microheterogeneity among structural proteins, and that this microheterogeneity is necessary for the elaboration of an unusually complex organelle characterized by a well-defined shape and dynamic properties.

MATERIALS AND METHODS

Cells and Culture Conditions

The wild-type Paramecium cells used in these experiments were *Paramecium tetraurelia* strain d4–2 and *Paramecium primaurelia* strain 156 (Sonneborn, 1974). Cells were grown at 27°C in wheat grass powder (Pines International, Lawrence, KA), inoculated with *Enterobacter aerogenes*, and supplemented with 0.4 μg/ml β-sitosterol (Sonneborn, 1970).

Two Dimensional (2D) Protein Gels

Trichocyst matrices were isolated as described by Sperling et al. (1987). The gel system was the same as that used to separate polypeptides for microsequencing (Le Caer *et al.*, 1990). Briefly, 2D protein gel electrophoresis was performed according to the method of Garrels (1979). In the first dimension, isoeletric focusing was carried out in the presence of a chemical spacer (50 mM MOPS) as described by Tindall (1986), at 500 V for 6 h using a Hoefer mini-tube gel system. The second dimension was a 13% sodium dodecyl sulfate-polyacrylamide gel electrophoresis.

N-Terminal Protein Sequences

The N-terminal microsequences of T1, T2, and T4 (with regions used for primer design underlined) are as follows (Tindall *et al.*, 1989; Le Caer *et al.*, 1990; Peterson *et al.*, 1990):

T1- FADQGALRDIVVAFNNLRVELVDSLNQITADEAEQVA
T2- DPLDRLLQTLTDLEDRYVAEQKEDDAKNQ
T4- GPVGEIQILLNNIASQLNGDQKKADKV

Genomic DNA Preparation

P. tetraurelia and *P. primaurelia* genomic DNA were isolated from exponential phase cultures as described by Dupuis (1992).

Polymerase Chain Reaction (PCR) Amplification and Sequence Analysis of Trichocyst Precursor Gene Fragments

Direct PCR. For direct PCR amplifications, partially degenerate oligonucleotide primers were designed on the basis of the N-terminal protein sequences given above. The nucleotide sequences, incorporating either an *Xba*I (sense primers) or an *Eco*RI (antisense primers) restriction enzyme site (underlined), were as follows:

T1: 5′- GCTCTAGATTYGCHGAYYARGGWGC - 3′
5′- GCGAATTCTCDGCYTCRTCDGCWGT - 3′
T2: a 5′- GCTCTAGAGAYCCWYTDGAYMG - 3′
b 5′- GCGAATTCTCYTCYTTYTRYTCDGC - 3′
c 5′- GCGAATTCTCYYTRRTTYTTDGCRTC -3′
T4: 5′- GCTCTAGAGGWCCWGTHGGWGA - 3′
5′- GCGAATTCACYTTRTCDGCYTTYTT - 3′

Two consecutive nested PCRs were performed to amplify T2 DNA sequences: primers T2a and T2c were used in the first reaction, a fraction of which (1–2 μl) was then used as template for the second amplification, which was carried out with primers T2a and T2b.

PCR reactions (50 μl) contained 50–200 pmol of each primer, 50 ng of genomic DNA, 0.1 mM dNTPs, and 2 U of *Taq* DNA polymerase (Boehringer, Mannheim, Germany). Reactions were overlaid with 50 μl of mineral oil (Sigma, St. Louis, MO) and carried through five cycles of denaturation at 90°C for 30 s, annealing at 40°C for 45 s, and extension at 72°C for 1 min 30 s, followed by 25 cycles of denaturation at 90°C for 30 s, annealing at 48°C for 45 s, and extension at 72°C for 1 min 30 s. After fractionation on agarose gels, the amplification products were recovered using the QIAEX Gel Extraction Kit (QIAGEN, Dusseldorf, Germany) and cloned into the *Xba*I and *Eco*RI sites of the pUC18 plasmid. DNA sequences were determined by the dideoxy nucleotide chain termination method using the T7 sequencing kit (Pharmacia, Uppsala, Sweden).

Inverse PCR. To amplify the DNA regions immediately flanking the gene fragments identified as described above, internal, outward-facing oligonucleotides were designed on the basis of the sequences determined by direct PCR and used to prime inverse PCRs on templates consisting of circularized genomic DNA fragments (Ochman *et al.*, 1990). The oligonucleotide sequences were as follows:

T1: sense 5′- GTTGATTCACTCAAYTAAATCAC - 3′
antisense 5′- CAACTCTCAARTTATTGAAGGC - 3′
T2: sense 5′ -TGGAAGACAGATATGTTG - 3′
antisense 5′- GTCRGTCAAGGTTTAGAG -3′
T4: sense 5′- GCCTCATAATTGAATGGAG - 3′
antisense 5′- AAAAGGATTTAGATCTCACC - 3′

Genomic DNA was digested by restriction enzymes (*Hin*dIII for T1 and T4, and *Alu*I for T2), diluted to a final concentration of ~1 ng/μl and then incubated for 16–18 h at 14°C in the presence of T4 DNA ligase (0.02 U/μl; Life Technologies, Bethesda, MD), conditions that favor the generation of monomeric DNA circles. PCR amplification was then carried out on circularized DNA molecules (DNA final concentration: 0.2–0.4 ng/μl) as described above. After cloning into the *Sma*I site of pUC18 according to standard protocols (Sambrook *et al.*, 1989), the inverse PCR products were sequenced, and then used to generate ^{32}P-labeled probes.

Preparation of Radioactive Probes

T1 intron probes were prepared using the following oligonucleotides, which correspond to two different specific intron sequences:

T1-c 5′- GTAATGTCTAATTGATATATCCTCTAG -3′
T1-f 5′- GTATTCTATTTTCTCATTCCTAG - 3′

Each oligonucleotide (1–2 pmol) was incubated for 1 h at 37°C with 10 μCi [γ-^{32}P]ATP (Amersham, Buckinghamshire, UK) in the presence of T4 kinase (Life Technologies; final concentration: 1 U/μl).

Exon probes were prepared by amplification of the cloned DNA fragments generated by inverse PCR. Each PCR reaction was carried

out with 10–30 ng of recombinant pUC18 plasmid containing the appropriate insert DNA, 6 mM dATP, 50 mM dCTP, 50 mM dGTP, 50 mM dTTP, in the presence of [α-^{32}P]ATP (Amersham; final concentration: 1 μCi/μl). For T1 and T2 probes (290 and 400 bp, respectively), the primers used were the same as those described above for the original inverse PCR amplification. New primers were designed to generate a 287-bp probe from the 1409-bp T4 DNA fragment obtained by inverse PCR:

sense 5′- CAGTTGCCACAGCTAGA - 3′
antisense 5′- AAGTTGGACAAGGCGAC - 3′

Southern Blots

Southern blot experiments were carried out using as probes either exact sequence oligonucleotides (T1 introns: Figure 5, lanes b and c) or DNA fragments generated by PCR as described above (Figure 5, lane a; Figures 6 and 7).

Two to five micrograms of genomic DNA were used for each restriction digestion, separated by electrophoresis on 0.5–1% agarose gels, and transferred to Hybond-N$^+$ membranes (Amersham) in 0.4 M NaOH.

Hybridizations were carried out as described (Church and Gilbert, 1984), either at 42°C (T1 intron probes) or at 60–62°C (PCR-generated exon probes). Filters were then washed with decreasing concentrations of SSC: 2× SSC for 20–30 min, followed by 0.2× SSC for 20–45 min, at the same temperatures utilized during the hybridization step (1× SSC: 0.15 M NaCl, 0.015 M sodium citrate, pH 7.2). Autoradiograms were obtained by exposing the membranes to Kodak X-OMAT AR (Eastman Kodak, Rochester, NY) or Amersham Hyperfilm-MP films at −80°C with an intensifying screen.

RNA Preparation and Northern Blots

Total RNA was prepared as described by Chomczynski and Sacchi (1987), except that the cells were lysed by vortexing in the presence of glass beads.

For Northern blots, RNA (20 μg/lane) was fractionated on formaldehyde - 1.4% agarose gels (Sambrook *et al.*, 1989) and transferred to Hybond-C extra filters (Amersham) in 0.15 M Na acetate, according to the method of Reed and Mann (1985).

Hybridizations were carried out at 48°C in 6× SSC, 2× Denhardt's solution according to the method of Sambrook *et al.* (1989); the filters were then washed at the same temperature, as described for Southern blots.

CHEF Electrophoresis and Blots

For the analysis of chromosomal DNA, high molecular wt Paramecium DNA was prepared in agarose inserts and then separated by contour-clamped homogeneous electric field (CHEF) electrophoresis as described by Caron (1992).

For the Southern blot shown in Figure 7, kindly provided by Dr. Eric Meyer (Laboratoire de Génétique Moléculaire, Ecole Normale Supérieure, Paris, France), macronuclear chromosomal DNA was fractionated on a 1% agarose gel at 140 V for 20 h, with a commutation time of 50 s.

Characterization of T1 cDNA

RT-PCR was performed using total *Paramecium* RNA according to standard procedures. Briefly, the RNA (2 μg) was incubated in the presence of 50 mM Tris-HCl, pH 8.3, 60 mM NaCl, 6 mM $MgCl_2$, 10 mM dithiothreitol, 0.5 mM dNTPs, and 250 nmol oligo$(dT)_{15}$ in a reaction vol of 20 μl. The mixture was incubated at 65°C for 5 min, then at 39°C for 10 min before the addition of AMV-reverse transcriptase (Pharmacia; 0.5 U/μl); cDNA synthesis was carried out at 39°C for 90 min. PCR amplification was then carried out as described above with 1–2 μl of the cDNA, in the presence of 50 pmol of each PCR primer. The primers (see Figure 3B) were either the T1 direct PCR primers given above (primer set 1) or the following primers (set 2):

sense 5′- GAACACAAYGAWGCTATYGG - 3′
antisense 5′- GTTGATTCACTCAAYTAAATCAC - 3′

The amplification products were cloned and sequenced as described above.

RESULTS

A Family of Related Proteins

Although the trichocyst matrix is a true three-dimensional crystal at low resolution (~50 Å), it shows structural disorder at molecular spacings (Sperling *et al*, 1987) consistent with the fact that it is built up from a heterogeneous set of polypeptides. High resolution 2D gel electrophoresis reveals a complex pattern of small acidic polypeptides consisting of 30 major and as many as 100 distinct spots (Figure 1b; Tindall, 1986). That these molecules constitute a family of related proteins is supported by several lines of evidence, beyond the fact that they assemble together into a periodic structure. First, the proteins are of similar size (15–20 kDa) and isoelectric point (pH 4.7–5.5). Second, the 15- to 20-kDa polypeptides of the mature trichocyst matrix are produced by proteolytic processing of a set of 40- to 45-kDa precursor molecules (Adoutte *et al.*, 1984; Gautier *et al.*, 1994). Finally, the proteins are immunologically related. For example, polyclonal antibodies raised against a small subset of the proteins recognize most or all of them (Adoutte *et al.*, 1984).

How does *Paramecium* generate such a large number of related proteins? Post-translational modifications could generate great complexity starting from one or a few precursor proteins, although, other than the proteolytic processing, post-translational modifications of trichocyst proteins have yet to be identified. Alternatively, the heterogeneity could result from coexpression of many genes coding for similar proteins. The first indication that this latter possibility might indeed be the case came from microsequencing of matrix polypeptides: nine gel spots chosen at random gave eight distinct N-terminal amino acid sequences (Le Caer *et al.*, 1990).

Multiple PCR Products Generated for Each of Three Mature Trichocyst Matrix Polypeptides

We have used N-terminal sequences determined for three of the 2D gel spots (T1, T2, and T4 in Figure 1, b and c) to design partially degenerate oligonucleotide primers for PCR amplification of the corresponding genomic sequences. *Paramecium* genomic DNA was amplified, the amplification products were cloned in a plasmid vector, and a number of clones were chosen for DNA sequencing.

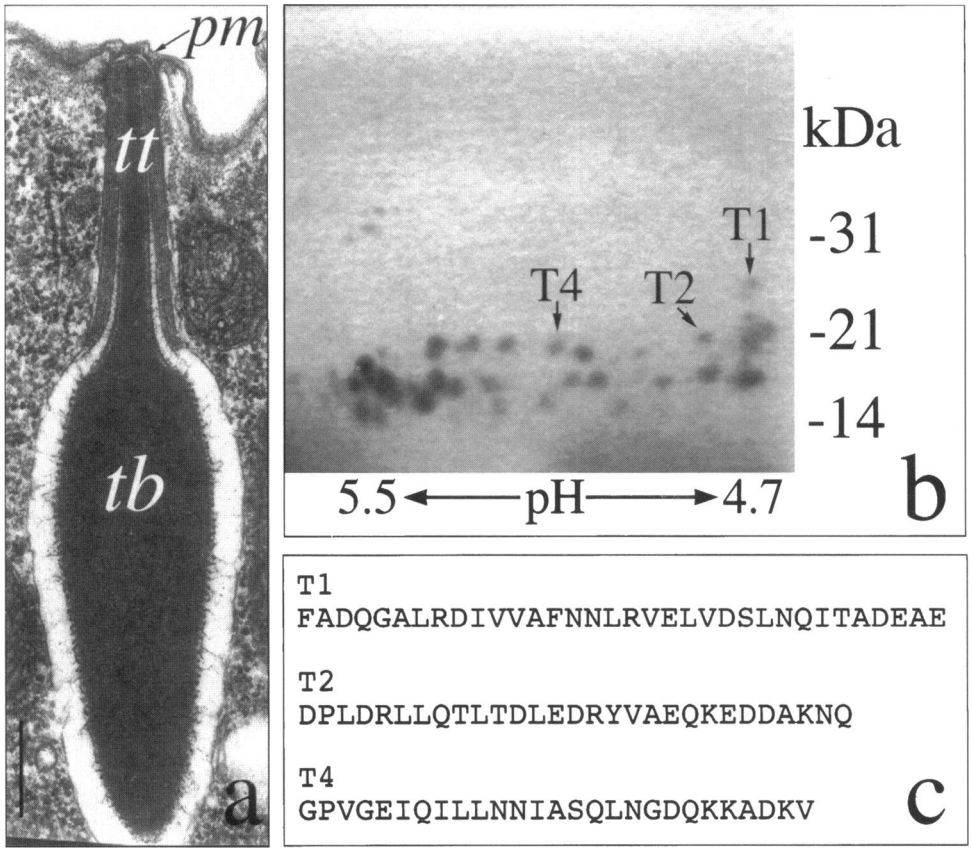

Figure 1. The crystalline content of *Paramecium* trichocysts is built up from a heterogeneous set of polypeptides. (a) An electron micrograph, courtesy of Nicole Garreau de Loubresse, showing a trichocyst docked at the plasma membrane (pm). Note the carrot-shaped crystalline body (tb) and the elaborate tip assembly (tt). Bar, 0.5 μm. (b) 2D gel electrophoretogram of purified trichocyst crystalline contents. (c) N-terminal amino acid sequences for the spots labeled T1, T2, and T4.

Figure 2 shows that for each N-terminal protein sequence, three to six different nucleotide sequences were obtained. Most of the differences can be attributed to silent substitutions that generate synonymous codons. If we discount the codons corresponding to the PCR primers, which can introduce base substitutions, 16 of 22 T1 codons analyzed are variable among the different sequences and five of these are sites of (conservative) amino acid substitution. Similarly, 8 of 13 T2 codons and 11 of 16 T4 codons are variable, although only two of the T2 codons and none of the T4 codons are sites of amino acid variation. Interestingly, one of these amino acid substitutions has been detected by microsequencing: the T2-a sequence (-L-S) was reported by Tindall *et al.* (1989), whereas we found the sequence T2-c (-L-Q).

The results presented in Figure 2 are supported by independent PCR experiments, especially for T1. By inverse PCR (see MATERIALS AND METHODS) we found sequences identical (except for the nucleotides corresponding to the amplification primers) to T1-b, T1-c, T1-e, T1-f, T2-c, and T4-a. Moreover, T1-a, -e, and -f were found by direct PCR with primers used for cDNA amplification experiments (see below). Finally, one member of each subfamily has been cloned from a genomic library (Gautier and Madeddu, unpublished data) and the gene sequences in the regions explored by the PCR experiments are identical to the T1-b, T2-c, and T4-a sequences. We thus consider it unlikely that the sequence variations we report result from PCR artifact.

The T1 genes contain typical *Paramecium* introns (Dupuis, 1992) inserted between two codons. These intron sequences, with the exception of the conserved 5′ and 3′ splice sites, present greater sequence variation than the surrounding coding sequences. Although there is little doubt that these elements are introns given the comparison of the nucleotide and amino acid sequences and their resemblance to known *Paramecium* introns (Russell *et al.*, 1994), experimental confirmation was obtained by RT-PCR using total *Paramecium* RNA as substrate. Amplified cDNA, which should no longer contain the intron, is indeed around 25-bp smaller than the amplified genomic DNA (Figure 3).

As only gene fragments and not complete genes have been sequenced for the present study, we verified the length of the complete transcripts by Northern blot experiments (Figure 4). T1, T2, and T4 probes all hybridize with RNA of the size (1.4–1.5 kbases) ex-

pected to code for the 40- to 45-KDa trichocyst precursor proteins.

Genomic Blots Probed with Exon and Intron Sequences Confirm that Several Genes Code for Nearly Identical Polypeptides

According to the PCR experiments, multiple DNA sequences correspond to each N-terminal microsequence. To obtain complementary evidence that several different genes code for each matrix polypeptide, we performed Southern blot experiments using *Paramecium* genomic DNA. To generate larger (~300 nucleotide) exon probes, corresponding in each case to a single gene sequence (T1-b, T2-c, and T4-a in Figure 2), the nucleotide sequences determined by direct PCR were used to amplify inverse PCR products that were cloned and sequenced, as detailed in MATERIALS AND METHODS. T1 intron probes consisted of synthetic oligonucleotides corresponding to specific intron sequences.

Figure 5 shows hybridization of genomic DNA from *P. tetraurelia* and from a related species, *P. primaurelia*. The T1 exon probe hybridizes with seven to nine different restriction fragments for both digests (*Eco*RI and *Hin*dIII), on the DNA of both *P. tetraurelia* and *P. primaurelia*. The T1 exon probe contains neither *Hin*dIII nor *Eco*RI sites, but we cannot exclude the possibility that some of the other T1 genes do contain sites for these enzymes in the region covered by the probe. We can exclude the possibility that the numerous bands detected result from partial digestions: all blots used in the present study were controlled by hybridization with a probe for the single copy *Paramecium* calmodulin gene (Kink *et al.*, 1991).

The same blots were hybridized with two different introns from the *P. tetraurelia* T1 genes (T1-c and T1-f introns; Figure 5, b and c, respectively). The intron probes only hybridize with the *P. tetraurelia* DNA, although *P. primaurelia* T1 genes do contain introns at the same position, as determined by sequencing products of a PCR experiment (Madeddu, unpublished results). The T1-c intron hybridizes with two of the genomic fragments, although with different intensities, and the T1-f intron with a single genomic fragment.

These Southern blot experiments show the following: 1) that approximately eight genes code for T1 proteins in the *Paramecium* macronuclear genome; 2) that the T1 coding sequences are highly conserved between *P. tetraurelia* and *P. primaurelia*, species that separated around 10 million years ago (H. Philippe and A. Baroin, personal communication); and 3) that the intron sequences have diverged more rapidly than the coding sequences and constitute gene-specific markers.

Similar hybridization experiments were carried out using probes for T2 and T4 gene sequences. As shown in Figure 6 for restriction digests with three different enzymes, both probes recognize a number of fragments, consistent with the existence of approximately eight genes in the macronuclear genome for both T2 and T4. The fact that the initial

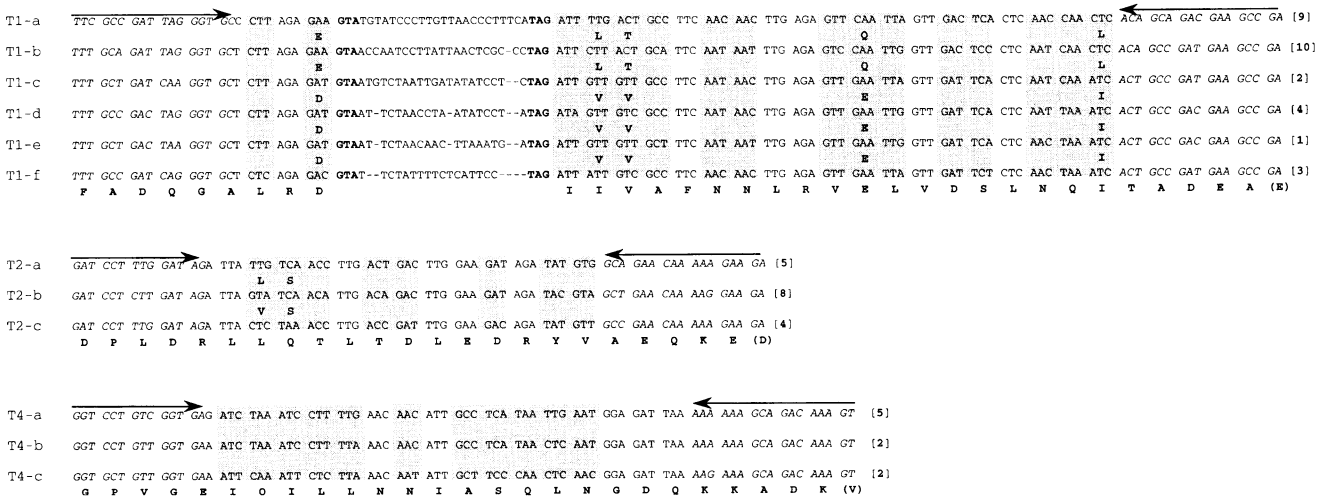

Figure 2. Each N-terminal sequence leads to the generation of multiple PCR products. *Paramecium* genomic DNA was amplified with partially degenerate oligonucleotide primers corresponding to N-terminal protein sequences, and the PCR products were cloned and sequenced. The different sequences obtained with each set of primers are represented. Arrows mark the extremities of the sequences, corresponding to the degenerate primers. Codons that vary among the different members of each group are shaded and amino acid changes are indicated underneath the codons. The conserved nucleotides of the 5′ and 3′ splice sites of the introns, which interrupt T1 sequences, are shown in boldface print. The number of times each sequence was found is given in square brackets.

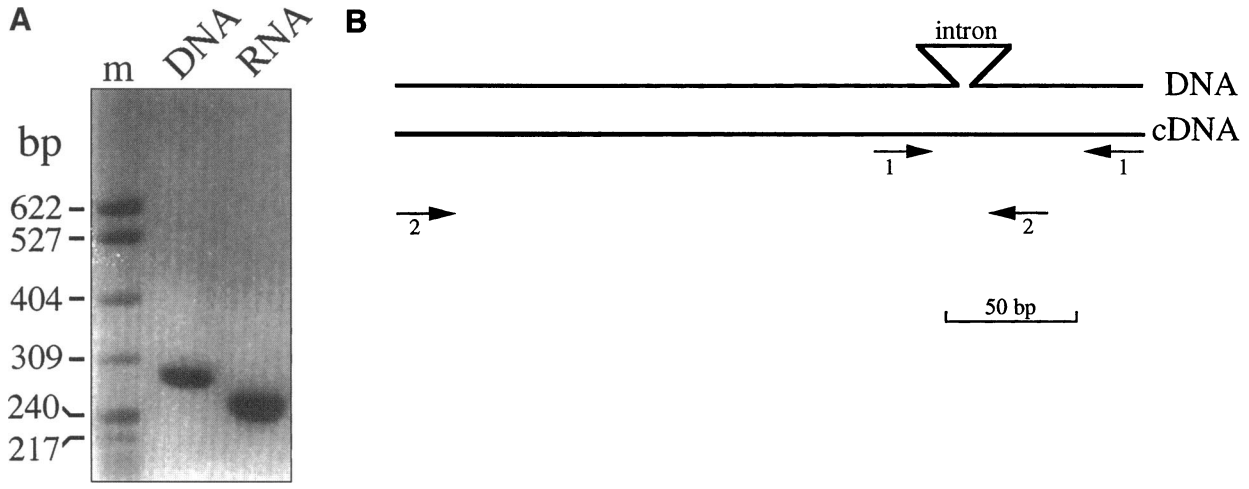

Figure 3. Presence of introns in T1 genes confirmed by RT-PCR. (A) T1 DNA fragments were generated by PCR amplification with primer set 2, using as template either genomic DNA or cDNA generated by reverse transcription of total RNA. The products were separated on a 3% NuSieve agarose gel and visualized by ethidium bromide staining. The cDNA amplification product appears to be smaller than the genomic DNA amplification product by approximately 25 bp, the average size of T1 introns. (B) Schematic representation of the positions of the two primer sets used to amplify T1 cDNA.

PCR experiments amplified only three distinct nucleotide sequences for T2 and for T4 may reflect a bias in the PCR experiments: a variable amino acid (or a rare codon) in the region used to design the partially degenerate PCR primers could lead to 3′ mismatches and failure to prime DNA synthesis. It is also possible that different genes have identical sequences in the region explored by PCR. As for the T1 blots, we cannot exclude the possibility that some members of each subfamily contain sites for these enzymes. However hybridization patterns with around eight bands were also obtained using several other restriction enzymes. A very conservative estimate would be that there are "only" four genes in the T2 and T4 subfamilies, corresponding to the most strongly hybridizing restriction fragments.

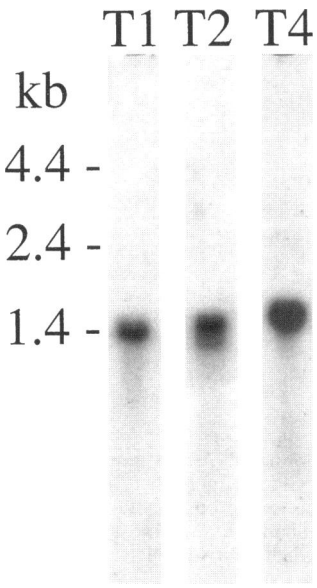

Figure 4. Characterization of transcript size by Northern blot. Total *P. tetraurelia* RNA was fractionated on denaturing gels, transferred to nitrocellulose membranes, and hybridized with the same T1, T2, and T4 exon probes used for the Southern blot experiments. RNA sizes were determined with a kit of RNA markers purchased from Life Technologies.

T1 Genes Are Coexpressed

The presence of introns in T1 genes provides a means of evaluating the expression of the different members of this subfamily. Two sets of primers that amplify fragments spanning the intron were used to amplify cDNA (Figure 3B). The absence of the intron sequences in the cloned amplification products constitutes proof that cDNA and not contaminating genomic DNA has been amplified. Table 1 summarizes the results of these experiments by showing how many times each T1 sequence was found among cDNA and genomic DNA amplification products for each set of primers.

With the original degenerate primers that correspond to the T1 N-terminal microsequence, cDNA amplification allowed us to find five of the six sequences shown in Figure 2, as well as two additional sequences that are presented in the Table legend. We cannot be certain that these sequences are bona fide new members of the T1 subfamily because we cannot evaluate the intron sequences and because they were found among the products of a single amplification reaction. The second set of primers are nearly sequence specific (see MATERIALS AND METHODS). Three different T1 sequences were found with these

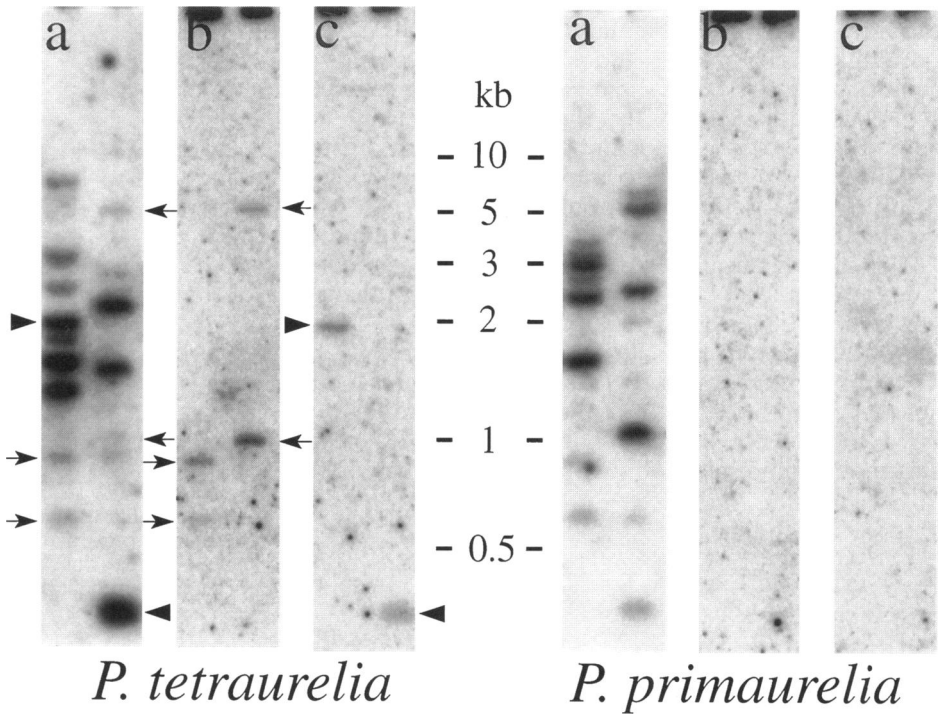

Figure 5. Genomic blots of T1 with intron and exon probes. *P. tetraurelia* and *primaurelia* genomic DNA were digested with *Eco*RI (left lanes) and *Hin*dIII (right lines), separated on a 1% agarose gel, and transferred to nylon membranes. The same blot was probed with (a) a 290-bp DNA fragment generated by PCR (exon probe), (b) an oligonucleotide probe corresponding to the intron T1-c, and (c) an oligonucleotide probe corresponding to the intron T1-f. Arrows indicate hybridization to the same restriction fragments. We note that a hybridization pattern equivalent to that in panel a is found if the mixture of first generation PCR products (obtained with the T1 degenerate primers) is used as probe, after labeling with T4 polynucleotide kinase. The results of a PCR experiment using *P. primaurelia* DNA as template and the T1 degenerate amplification primers corroborate the genomic blot data. Fifteen clones were sequenced and gave five different sequences, with introns at the same positions as found in the *P. tetraurelia* sequences. The coding regions most closely resemble those of T1-a, -b, -c, -d, and -f: there are no amino acid differences and from two to eight silent base substitutions between homologous sequences.

primers when genomic DNA was used as template. Although only five cloned cDNA amplification products were analyzed, the same three sequences were found, including the T1-f sequence not found with the first set of primers.

Because all characterized T1 genomic DNA sequences were found among the cDNA amplification products at least once, we conclude that all macronuclear T1 genes are expressed. By extrapolation to the other subfamilies, we consider it likely that all the different members of the trichocyst protein multigene family are coexpressed.

Each Subfamily Has a Distinct Location on Paramecium Macronuclear Chromosomes

To see whether the genes coding for trichocyst proteins are physically linked in the macronuclear genome, chromosomes were separated by CHEF gel electrophoresis and the corresponding blot was hybridized consecutively with T1, T2, and T4 gene probes. Each probe hybridizes with only two bands (possibly three for T4), far fewer than the number of bands on the Southern blots (Figures 5 and 6). If each band corresponds to a single macronuclear chromosome, then at least some of the genes belonging to each subfamily are physically linked in the macronuclear genome. The three subfamilies however are not linked to each other, because the T1, T2, and T4 probes hybridize to different sets of chromosomes. Each subfamily has a distinct location in the macronucleus.

DISCUSSION

100 Genes to Construct a Crystal

The crystalline core of *Paramecium* secretory granules is built up from a heterogeneous set of immunologically related acidic polypeptides. N-terminal microsequencing of a number of these polypeptides, by three groups using different approaches, has yielded a total of nine different amino acid sequences (reviewed in Madeddu *et al.*, 1994). None of the approaches sought to systematically microsequence all of the different polypeptides, so that it is likely that many more distinct N-terminal sequences exist among this complex set of molecules.

Starting from three of the available N-terminal sequences, we have amplified genomic DNA and obtained evidence both from the sequences of the amplification products and from Southern blot experiments, that at least four and probably eight different genes code for nearly identical versions of each of the mature polypeptides. *Paramecium* thus appears to use a large

multigene family to code for the proteins stored in its secretory granules.

The total number of genes coding for trichocyst matrix proteins can be tentatively estimated. The 2D protein map (Figure 1b; Tindall *et al.*, 1989) consists of some 30 major spots. If the major spots have different sequences, as suggested by the microsequencing results (Le Caer *et al.*, 1990), and given that each precursor probably gives rise to two mature polypeptides (Gautier and Madeddu, unpublished data), we can postulate the existence of 15 different gene subfamilies. If each subfamily, like the three we have studied, consists of at least four and in some cases eight genes, there should be around 100 genes in all.

We have shown that all identified members of the T1 family are expressed. Because there is no reason to suspect different behavior of the other subfamilies, we have (to our knowledge) the first example of the constitutive coexpression of such a large number of related genes. The correspondence between these genes and the 2D protein map is not yet clear. For example, do T1 gene products all map to the same spot, or do they account for eight of the spots? Future experiments will address this question.

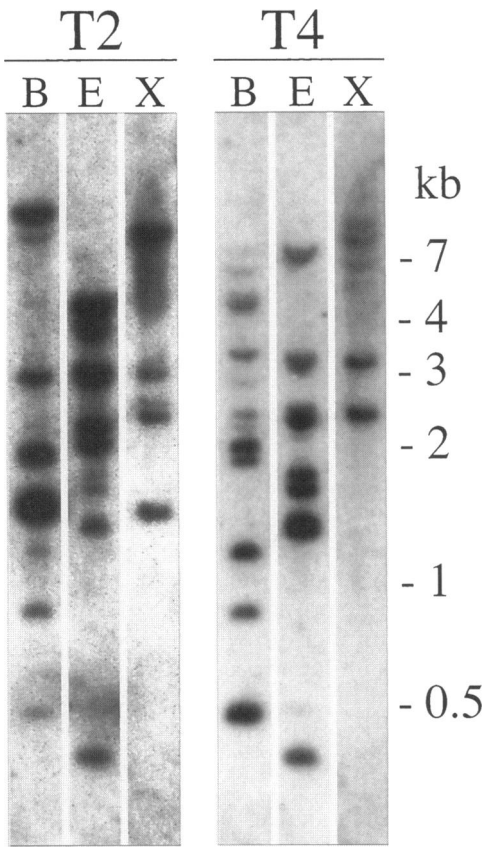

Figure 6. Genomic blots with T2 and T4 probes. *P. tetraurelia* genomic DNA was digested with *Bgl*II (B), *Eco*RI (E), and *Xba*I (X), separated on an agarose gel and transferred to a nylon membrane. The same blot was hybridized with a 400-bp T2 probe and a 287-bp T4 probe. Equivalent hybridization patterns are found if the mixture of first generation PCR products (obtained with the degenerate primers) are labeled with T4 polynucleotide kinase and used as a probe.

Table 1. cDNA amplification indicates that all T1 genes are expressed

	1st set of primers		2nd set of primers	
	DNA	cDNA	DNA	cDNA
T1-a	9	3	6	3
T1-b	10	1		
T1-c	2	4		
T1-d	4	6		
T1-e	1	1	3	1
T1-f	3		4	1
T1-g		3		
T1-h		2		

The table shows the number of different PCR clones found for each T1 sequence in PCR experiments using either genomic DNA or cDNA as template. Independent experiments were carried out with two sets of primers that span T1 introns; the relative positions of the primers with respect to a T1 gene are shown in Figure 3B. The six T1 gene sequences originally identified by PCR (minus the introns) as well as two new sequences (T1-g and T1-h) were found among the cDNA amplification products. The gene sequences for T1-a to -f are given in Figure 2, and the two additional cDNA sequences are given below (the parts of the sequences corresponding to the degenerate primers are in italics).

T1-g 5′*TTC GCC GAT CAG GGT GC*T CTT AAA GAT ATC CTC GTT GCA TTC AAC AAT TTG AGA GTT TAA TTA GTT GAT TCA CTT AAC ACA ATC *ACA GCC GAT GAA GCA GA* 3′

T1-h 5′*TTT GCA GAC TAG GGT GCT* CTT AGA GAT ATT GTT GTT GCC TTC AAT AAC TTG AGA GTT GAA TTA GTT GAC TCC CTC AAT TAA ATC *ACT GCA GAC GAA GCA GA* 3′

A Large Multigene Family in a Unicellular Organism

The size of multigene families generally increases in the course of evolution, and the increase in complexity corresponds to developmental and tissue-specific differentiation of the functions and/or the expression of the different members of the family. In both free-living and parasitic protists, multigene families code for variable surface antigens, only one of which is expressed at a time (Sonneborn, 1948; Borst, 1986; Caron and Meyer, 1989). However these families consist of only 10–20 genes. We are aware of no other examples of a multigene family (rDNA and histone gene clusters excepted) as large as that reported here in a unicellular organism.

In *Paramecium*, the trichocyst multigene family is probably not the only family of coexpressed genes coding for a set of very similar proteins. Other ele-

ments of cellular architecture are also built up from families of related polypeptides. The ciliary rootlets, which assure the antero-posterior alignment of basal bodies and cilia, are composed of at least five related polypeptides (Sperling *et al.*, 1991). The epiplasm, a fibrous meshwork that lies beneath the plasma membrane, is composed of dozens of related polypeptides (Nahon *et al.*, 1993). The infraciliary lattice, a contractile network located at the interface of the cortex and the cytoplasm, consists of seven major related polypeptides that are most likely *Paramecium* homologues of centrin (Garreau de Loubresse *et al.*, 1991; and Klotz, Madeddu, Le Caer, Garreau de Loubresse, Ruiz, Salisbury, and Beisson, unpublished data).

The question arises as to how *Paramecium* generates these multigene families. Any approach to this question must take into account the nuclear dimorphism of ciliates: the coexistence in a single cytoplasm of a transcriptionally inactive diploid micronucleus (germ line) and a transcriptionally active polyploid macronucleus (somatic line), the macronuclear chromosomes being derived from the micronuclear chromosomes by a series of rearrangements that include amplification, fragmentation, internal sequence elimination, and telomerization (Blackburn and Karrer, 1986). Although we fully expect to find all of the same trichocyst gene sequences in the micronucleus as in the macronucleus, we must demonstrate it, so as to exclude the possibility that new versions of these genes are created during the macronuclear differentiation process as a result of a novel editing mechanism.

Hybridization experiments with macronuclear chromosomes show that the three different subfamilies we have characterized are located on different macronuclear chromosomes. The fact that few chromosomes hybridize with the probes for each subfamily suggests that the genes of each subfamily may be physically linked in the micronucleus, given the polymorphism of *Paramecium* macronuclear chromosomes (Forney and Blackburn, 1988; Caron, 1992; Amar, 1994). We can postulate that a first and quite ancient series of gene duplications, followed by differentiation to give proteins with somewhat different sequences, gave rise to the subfamilies that are now distributed on different chromosomes. A second, more recent series of duplications would have populated the subfamilies with genes coding for nearly identical proteins, and these genes are still physically linked. It will be interesting to determine whether other species of ciliates such as *Pseudomicrothorax dubius*, whose trichocyst proteins immunologically cross-react with those of *P. tetraurelia* (Eperon *et al.*, 1993), use homologous genes to code for their secretory granule proteins, and to see whether they constitute a large multigene family.

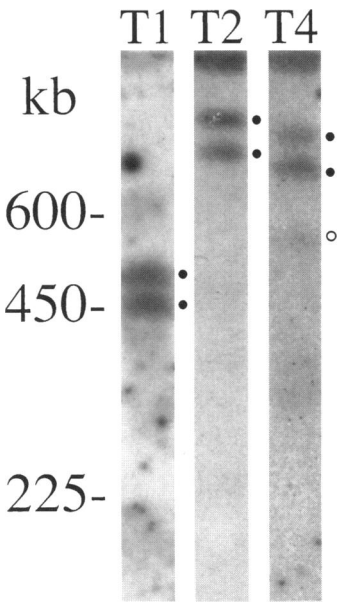

Figure 7. Macronuclear chromosome blots. *P. tetraurelia* chromosomal DNA was separated by CHEF electrophoresis, transferred to a nylon membrane, and probed with the T1, T2, and T4 exon probes. Yeast chromosomes were used as size markers. Hybridizing bands are marked by filled circles. A band hybridizing weakly with the T4 probe (which could result from cross-hybridization, a less abundant chromosome, or a chromosome bearing fewer T4 genes than the others) is marked with an open circle.

Redundancy or Functional Microheterogeneity?

The trichocyst matrix has a crystalline structure with periodicities in three dimensions, at least at low resolution (Sperling et al., 1987). One way in which distinct polypeptides could assemble into a periodic structure is if they are more or less interchangeable, sharing for example, common secondary or even tertiary structure. This is in fact compatible with the sequence data currently available that does indicate high α-helical content and probable coiled coil domains in these proteins (Gautier and Madeddu, unpublished data). Protein microheterogeneity could be instrumental in the generation of shape if different polypeptides integrate the crystalline edifice at distinct times in the assembly process, perhaps as a function of the continually changing concentration of a counter ion.

An alternative hypothesis is that the size of the multigene family is a means of assuring adequate protein synthesis. The differences in sequence would then reflect the accumulation of neutral mutations, compatible with the molecular design (coiled-coil?) needed for matrix assembly. One argument against redundancy of the genes and their products is the fact that the polyploidy (800n) of the macronuclear genome should be sufficient to assure the production of large amounts of proteins. For example, the surface antigen, which coats the entire cell surface including the cilia, amounts to ~3% of total cellular protein yet results from the expression of a single surface antigen gene (Caron and Meyer, 1989).

We envisage an experimental approach, involving transformation experiments and specific anti-peptide antibodies, to test these alternative hypotheses as to the selective pressure that has favored the evolution of the trichocyst multigene family. In the meantime, it is

tempting to suggest that the trichocyst multigene family has arisen in *Paramecium* to provide a microheterogeneity of structural proteins necessary for the design and assembly of a geometrically and dynamically complex crystalline edifice. Trichocyst morphogenesis may constitute an extreme example, given the number of different genes used, of a general strategy employed by eukaryotic cells for the construction of dynamic subcellular structures. A different way to generate microheterogeneity is exemplified by the microtubule-based cytoskeleton. A limited number of structural genes, in conjunction with post-translational modifications such as polyglutamylation (Eddé *et al.*, 1990) and polyglycylation (Redeker *et al.*, 1994) capable of generating a great number of isoforms, also leads to considerable microheterogeneity among proteins that co-assemble. The use of a large multigene family to generate microheterogeneity puts the blueprint in the genome where it is not easily modified whereas use of post-translational modifications permits architectural plasticity, in the course of development or in response to the environment.

ACKNOWLEDGMENTS

We thank L. Amar and P. Dupuis for critical reading of the manuscript and J. Cohen and A. Adoutte for many useful discussions. We are indebted to E. Meyer for help with the chromosome blot experiments. We are particularly grateful to J. Beisson for support, encouragement, and critical advice throughout the present work. L.M. was supported by a senior fellowship of the EEC Bridge Program. M.-C.G. was supported by a graduate fellowship from the Ministère de l'Enseignement Supérieur et de la Recherche (MESR). This work was financed by a contract from the Genome Program of the MESR (GIP GREG) and by the CNRS.

Note added in proof: The GenBank accession numbers for the partial coding sequences presented in Figure 2 are: T1a, U27503; T1b, U27504; T1c, U27505; T1d, U27506; T1e, U27507; T1f, U27508; T2a, U27509; T2b, U27510; T2c, U27511; T4a, U27512; T4b, U27513; and T4c, U27514.

REFERENCES

Adoutte, A. (1988). Exocytosis: biogenesis, transport and secretion of trichocysts. In: *Paramecium*, ed. H.-D. Görz, Heidelberg, Germany: Springer-Verlag, 325–362.

Adoutte, A., Garreau de Loubresse, N., and Beisson, J. (1984). Proteolytic cleavage and maturation of the crystalline secretion products of *Paramecium*. J. Mol. Biol. *180*, 1065–1080.

Amar, L. (1994). Chromosome end formation and internal sequence elimination as alternative genomic rearrangements in the ciliate *Paramecium*. J. Mol. Biol. *236*, 421–426.

Black, S.D., and Coon, M.J. (1987). P-450 cytochromes: structure and function. Adv. Enzymol. *60*, 35–87.

Blackburn, E.H., and Karrer, K.M. (1986). Genomic reorganization in ciliated protozoans. Annu. Rev. Genet. *20*, 501–521.

Borst, P. (1986). Discontinuous transcription and antigenic variation in trypanosomes. Annu. Rev. Biochem. *55*, 701–732.

Buck, L., and Axel, R. (1991). A novel multigene family encodes odorant receptors: a molecular basis for odor recognition. Cell *65*, 175–187.

Caron, F. (1992). A high degree of macronuclear chromosome polymorphism is generated by variable DNA rearrangements in *Paramecium primaurelia* during macronuclear differentiation. J. Mol. Biol. *225*, 661–678.

Caron, F., and Meyer, E. (1989). Molecular basis of surface antigen variation in *Paramecium*. Annu. Rev. Microbiol. *43*, 185–188.

Chomczynski, P., and Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal. Biochem. *162*, 156–159.

Church, G.M., and Gilbert, W. (1984). Genomic sequencing. Proc. Natl. Acad. Sci., USA *81*, 1991–1995.

Clark, A.G. (1994). Invasion and maintenance of a gene duplication. Proc. Natl. Acad. Sci, USA *91*, 2950–2954.

Dupuis, P. (1992). The β-tubulin genes of *Paramecium* are interrupted by two 27 bp introns. EMBO J. *11*, 3713–3719.

Eddé, B., Rossier, J., Le Caer, J.P., Desbruyères, E., Gros, F., and Denoulet, P. (1990). Posttranslational glutamylation of alpha-tubulin. Science *247*, 83–85.

Eperon, S., Vigues, B., and Peck, R.K. (1993). Immunological characterization of trichocyst proteins in the ciliate *Pseudomicrothorax dubius*. J. Eukaryot. Microbiol. *40*, 81–89.

Forney, J.D., and Blackburn, E.H. (1988). Developmentally controlled telomere addition in wild-type and mutant *Paramecia*. Mol. Cell. Biol. *8*, 251–258.

Garreau de Loubresse, N., Klotz, C., Vigues, B., Rutin, J., and Beisson, J. (1991). $Ca2^{++}$-binding proteins and contractility of the infraciliary lattice in *Paramecium*. Biol. Cell *71*, 217–225.

Garrels, J.I. (1979). Two-dimensional gel electrophoresis and computer analysis of proteins synthesized by clonal cell lines. J. Biol. Chem. *254*, 7961–7977.

Gautier, M.-C., Garreau de Loubresse, N., Madeddu, L., and Sperling, L. (1994). Evidence for defects in membrane traffic in *Paramecium* secretory mutants unable to produce functional storage granules. J. Cell Biol. *124*, 893–902.

Goldsmith, M.R., and Kafatos, F.C. (1984). Developmentally regulated genes in silkmoths. Annu. Rev. Genet. *18*, 443–487.

Kink, J.A., Maley, M.E., Ling, K.Y., Kanabrocki, J.A., and Kung, C. (1991). Efficient expression of the Paramecium calmodulin gene in *Escherichia coli* after four TAA-to-CAA changes through a series of polymerase chain reactions. J. Protozool. *38*, 441–447.

Le Caer, J.P., Rossier, J., and Sperling, L. (1990). Crystalline contents of *Paramecium* secretory vesicles: N-terminal sequences of a family of polypeptides separated on 2-D minigels and blotted onto immobilon. J. Prot. Chem. *9*, 290–291.

Madeddu, L., Gautier, M.C., Le Caer, J.P., Garreau de Loubresse, N., and Sperling, L. (1994). Protein processing and morphogenesis of secretory granules in *Paramecium*. Biochimie *76*, 329–335.

Nahon, P., Coffe, G., Le Guyader, H., Darmanaden-Delorme, J., Jeanmaire-Wolfe, R., Clérot, J.-C., and Adoutte, A. (1993). Identification of the epiplasmins, a new set of cortical proteins of the membrane cytoskeleton in *Paramecium*. J. Cell Sci. *104*, 975–990.

Ochman, H., Medhora, M.M., Garza, D., and Hartl, D.L. (1990). Amplification of flanking sequences by inverse PCR. In: PCR Protocols, ed. M.A. Innis, D.H. Gelfand, J.J. Sninsky, and T.J. White, San Diego, CA: Academic Press, 219–227.

Ohta, T. (1991). Multigene families and the evolution of complexity. J. Mol. Evol. *33*, 34–41.

Peterson, J.B., Nelson, D.L., and Angeletti, R.H. (1990). Relationships of *Paramecium* and endocrine secretory proteins. In: Current Research in Protein Chemistry, ed. J.J. Villafranca, New York: Academic Press, 79–85.

Redeker, V., Levilliers, N., Schmitter, J.M., Le Caer, J.P., Rossier, J., Adoutte, A., and Bré, M.H. (1994). Polyglycylation of tubulin: a posttranslational modification in axonemal microtubules. Science *266*, 1688–1691.

Reed, K., and Mann, D. (1985). Rapid transfer of DNA from agarose gel to nylon membrane. Nucleic Acids Res. *13*, 7207–7221.

Russell, C.B., Fraga, D., and Hinrichsen, R.D. (1994). Extremely short 20–33 nucleotide introns are the standard length in *Paramecium tetraurelia*. Nucleic Acids Res. *22*, 1221–1225.

Sambrook, J., Fritsch, E., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Sonneborn, T.M. (1948). The determination of hereditary antigenic differences in genically identical *Paramecium* cells. Proc. Natl. Acad. Sci. USA *34*, 413–418.

Sonneborn, T.M. (1970). Methods in *Paramecium* research. Methods Cell. Physiol. *4*, 241–339.

Sonneborn, T.M. (1974). *Paramecium aurelia*. Handbook of Genetics, ed. R. King, New York: Plenum Publishing, 469–594.

Sperling, L., Keryer, G., Ruiz, F., and Beisson, J. (1991). Cortical morphogenesis in *Paramecium*: a transcellular wave of protein phosphorylation involved in ciliary rootlet disassembly. Dev. Biol. *148*, 205–218.

Sperling, L., Tardieu, A., and Gulik-Krzywicki, T. (1987). The crystal lattice of *Paramecium* trichocysts before and after exocytosis by X-ray diffraction and freeze-fracture electron microscopy. J. Cell Biol. *105*, 1649–1662.

Tindall, S.H. (1986). Selection of chemical spacers to improve isoelectric focusing resolving power: implication for use in two-dimensional electrophoresis. Anal. Biochem. *159*, 287–294.

Tindall, S.H., De Vito, L.D., and Nelson, D.L. (1989). Biochemical characterization of the *Paramecium* secretory granules. J. Cell. Sci. *92*, 441–447.

Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature *302*, 575–581.

Wilson, D.R., and Larkins, B.A. (1984). Zein gene organization in maize and related grasses. J. Mol. Evol. *20*, 330–340.