The Trypanosoma cruzi Mucin Family Is Transcribed from Hundreds of Genes Having Hypervariable Regions*

Javier M. Di Noia§§, Iván D’Orso‡, Lena Åslund¶, Daniel O. Sánchez‡‡‡, and Alberto C.C. Frasch***‡‡‡

From the §Institute of Investigaciones Biotecnológicas, Universidad Nacional de General San Martín, Casilla de Correo 30, 1650, Provincia de Buenos Aires, Argentina and ¶Department of Medical Genetics, Biomedical Center, Uppsala University, Box 588, S-751 23, Uppsala, Sweden

In previous works we have identified genes in the protozoan parasite Trypanosoma cruzi whose structure resemble those of mammalian mucin genes. Indirect evidence suggested that these genes might encode the core protein of parasite mucins, glycoproteins that were proposed to be involved in the interaction with, and invasion of, mammalian host cells. We now show that the mucin gene family from T. cruzi is much larger and diverse than expected. A minimal number of 484 mucin genes per haploid genome is calculated for a parasite clone. Most, if not all, genes are transcribed, as deduced from cDNA analysis. Comparison of the cDNA sequences showed evidences of a high mutation rate in localized regions of the genes. Sequence conservation among members of the family is much higher in the untranslated (UTR) regions than in the sequences encoding the mature mucin core protein. Transcription units can be classified into two main subfamilies according to the sequence homologies in the 5′-UTR, whereas the 3′-UTR is highly conserved in all clones analyzed. The common origin of members of this gene family as well as their relationships can be defined by sequence comparison of different domains in the transcription units. The regions encoding the N and C termini, supposed to correspond to the leader peptide and membrane-anchoring signal, respectively, (Di Noia, J. M., Sánchez, D. O., and Frasch, A. C. C. (1995) J. Biol. Chem. 270, 24146–24149) are highly conserved. Conversely, the central regions are highly variable. These regions encode the target sites for O-glycosylation and are made of a variable number of repetitive units rich in Thr and Pro residues or are nonrepetitive but still rich in Thr/Ser and Pro residues. The region putatively coding for the N-terminal domain of the mature core protein is hypervariable, being different in most of the transcripts sequenced. Nonrepetitive central domains are unique to each gene. Gene-specific probes show that the relative abundance of different mRNAs varies greatly within the same parasite clone.

Mucins are highly glycosylated proteins in which the oligosaccharide chains are attached to the hydroxyl group of Thr and/or Ser residues. The core protein of most mucins is characterized by the presence of a variable central domain with repetitive amino acid units rich in Ser/Thr and Pro residues flanked by nonrepetitive N and C termini (1). Mucins have relevant roles in protection of epithelial cells and cell-cell interaction in higher organisms (2).

Trypanosoma cruzi, the unicellular parasite agent of Chagas disease, also expresses mucin-like molecules. Mucins are present at the parasite surface in all the developmental stages: epimastigote (3), metacyclic trypomastigote (4), amastigote (5), and cell-derived trypomastigote (6). They are linked to the surface membrane through a glycophosphatidylinositol anchor and are likely to be shed into the medium during the cell invasion process (4, 5). These parasite glycoproteins seem to have essential functions and unusual characteristics. T. cruzi mucins were suggested to be involved in the invasion of host cells by the parasite (7–9), an essential step for parasite multiplication and survival in mammals. Recently, T. cruzi mucins were shown to stimulate synthesis of interleukin 12 and tumor necrosis factor α cytokines by macrophages (5). One unusual structural feature of this parasite mucins is that the first sugar attached to the Ser/Thr residues is not N-acetylgalactosamine but N-acetylglicosaminic (3). Another rather surprising finding is the possible large number of different mucins in the parasite. Biochemical evidence suggested the existence of two groups of glycoproteins: those of apparent molecular mass 35–50 kDa (10) and others producing a large smear in Western blots from 80 to 200 kDa (6, 11). The recent identification of genes encoding the parasite core protein (12, 13) confirmed the existence of a gene family (14, 15), although the number of members has not been estimated yet. The first T. cruzi gene isolated was shown to encode a small protein of about 16 kDa having a repetitive central domain with the consensus sequence Thr-Lys-Pro2 flanked by nonrepetitive N and C termini (12). Characterization of a number of genes showed that, whereas flanking regions were conserved among strains, the central domain might vary greatly in the number of repeat units and that there...
is a second group of genes lacking repeats, with a variable central domain still rich in Thr and Pro and also in Ser residues. The multigene family was named TcMUC for \textit{T. cruzi} and in analogy with vertebrate mucin genes. Interestingly, some of these genes seem to be genetically linked to trans-sialidase (15), the enzyme that transfers sialic acid to the parasite mucins (11, 16). Although indirect, several evidences suggest that these genes indeed encode the core protein of parasite mucins as follows. 1) Deduced amino acid composition and structure is typical of apo-mucins (14). 2) Antibodies to a recombinant protein derived from a TcMUC gene and a monoclonal antibody directed against a carbohydrate epitope present in \textit{T. cruzi} mucins detected the same bands in Western blots (13). 3) When transfected into an eukaryotic cell, one TcMUC gene renders a product likely to be highly O-glycosylated (13). 4) It has been recently shown that a peptide derived from the sequence of TcMUC repetitive genes is a very good substrate for an UDP-N-acetylgalcosamine:peptide-N-acetylgalcosaminyltransferase activity thought to be the first enzyme involved in O-glycosylation machinery in \textit{T. cruzi}.1

In this work we have identified a number of different mRNAs transcribed from a minimal estimate of 484 mucin genes present in \textit{T. cruzi}. Although untranslated regions and some coding regions are highly conserved among members of this large family, definite coding regions are very variable; in particular, those encoding the mature N termini of the mucin core proteins.

**EXPERIMENTAL PROCEDURES**

**Parasites:** \textit{T. cruzi—}CL-Brener cloned stock (17) epimastigotes (a form of the parasite similar to the one present in the insect vector) were grown in brain-heart infusion medium (18). Cell-derived trypanotumastigotes (a form similar to that present in the mammalian bloodstream) were grown in Vero cell culture (19).

**Nucleic Acids Purification—**DNA was prepared from epimastigotes of \textit{T. cruzi} using conventional proteinase K phenol/chloroform method (20). RNA from the epimastigote and trypanotumastigote forms was prepared using TRIzol (Life Technologies, Inc.) according to manufacturer’s indications or by the CsCl/guanidinium isothiocyanate method (21).

**High Density Filter Hybridization—**A previously described genomic \textit{T. cruzi} cosmId library constructed in the LawrI7 vector and spotted on nitrocellulose membrane (22) was hybridized using the complete \textit{T. cruzi} genome as probe. The gene \textit{MUC.CA-2} was used as a positive control. The \textit{T. cruzi} cosmId library was screened using high stringency conditions were conducted onto high density filter arrays of the CL-Brener cosmid library generated using high stringency conditions for priming and hence some nonspecific cDNA synthesis, hemi-nested PCR was carried out. The first PCR reaction was done with a spliced leader primer named ME33 (5'-AACACATCTATTCA-3') and the P2 primer. Position of the second PCR primer internal to P2 position named STOP (5'-TGCAGCAGAGG- GTGTACG-CC-3') because it contained the stop codon of the genes, were used. PCR products were ligated into pGEM-T vector (Promega). Positive clones were identified by hybridization with \textit{MUC.CA-2} probe in high stringency conditions (0.1x SSC, 65 °C). Thirty clones obtained with primers ME33-STOP and ME33-P2 were sequenced.

**Isolation of \textit{T. cruzi} Mucin Transcripts—**A total of 59 clones containing total or partial mRNA corresponding to \textit{T. cruzi} mucin genes were obtained using the two strategies described above. From the library screening, 29 positive clones were isolated and partially or fully sequenced. In 13 of these clones, the cDNA spanned the region from the poly(A) tail to that encoding the signal peptide or even to the 5'-UTR, allowing them, the transcriptional acceptor site could be determined. The remaining 16 clones had the 3'-UTR, the region corresponding to the C terminus and a variable part of the central region.

For RT-PCR, we took advantage of the fact that all mRNA of trypanosomatids are matured at their 5’ end by trans-splicing, a process in which a 39 nt invariable sequence called spliced leader is attached to the 5’ end of the primary transcripts. From this strategy, 30 clones containing a complete ORF were analyzed, 13 corresponding to epimastigotes and 17 to trypanotumastigotes. All clones were named EMUC for expressed mucin gene, and EMUCe- and EMUCt- indicates if they were derived from epimastigote or trypanotumastigote stages, respectively. Those obtained from RT-PCR are identified by a number code (i.e. EMUCe-1, EMUCt-18), and those from library screening, by an alphanumeric code (i.e. EMUCe-24p20). As the “Results” section mainly deals with mucin family variability, transcripts are grouped without further specification of their source except when specified.

**DNA Sequencing—**Sequencing was performed using the Sequenase 2.0 kit (U. S. Biochemical Corp.) and an ABI 373 DNA sequencer (Perkin-Elmer). All sequence computational analyses were done using Lasergene package from DNASTAR INC.

**RESULTS**

\textit{T. cruzi Mucin Gene Family Is Composed by a Minimal Number of 484 Genes in One Parasite Clone—}Southern blot analysis of genomic DNA from \textit{T. cruzi} CL-Brener clone probed with a complete mucin gene showed a very complex pattern (not shown) similar to the one seen for all the other clones tested (14). This suggested a large number of genes as being part of this family. To estimate this number, hybridizations using high stringency conditions were conducted onto high density filter arrays of the CL-Brener cosmId library generated for mapping of the \textit{T. cruzi} genome (Ref. 22, see “Experimental Procedures”). There were 242 positive clones/haploid genome for a probe containing the complete mucin gene \textit{MUC.CA-2} (14). To estimate a number of mucin genes per cosmid clone, the EcoRI restriction patterns of 22 randomly chosen cosmId clones were analyzed by hybridization with the same probe (Fig. 1). None of the genes analyzed so far have an internal citrate. Washes were always at 65 °C in 0.1x SSC, 0.1% SDS.

\textit{cDNA Library Screening—}A normalized cDNA library constructed from epimastigotes of the CL-Brener clone and provided by Dr. E. Rondinelli (Rio de Janeiro, Brazil) had previously been ordered into 394-well microtiter plates. Replica filters of this cDNA library were hybridized using the complete \textit{MUC.CA-2} probe (14). The genes within the insert was PCR-amplified using the primers P1 and P2 and labeled with [α-32P]dCTP by random priming using MegaPrime (Amersham). The hybridization was performed in 0.5 M sodium phosphate buffer (pH 7.2) containing 7% SDS, 1 mM EDTA, and 100 μg/ml tRNA at 65 °C. The filters were washed in 40 mM sodium phosphate buffer (pH 7.2) at 65 °C for 2 h. The screening yielded 48 positive clones of different strengths in hybridization signals. Out of these, 29 were fully or partially sequenced.

**RT-PCR—**Reverse transcription was performed on total RNA from epimastigotes and cell-derived trypanotumastigotes using the Superscript kit following manufacturer instructions (Life Technologies, Inc.). Primer P2 (14) was used for cDNA synthesis designed on the conserved 3’ extragenic regions of available sequences and positioned at about 70 base pairs downstream of the stop codon. Due to the low temperature necessary to perform the reverse transcription reaction, that allowed a low stringency for priming and hence some nonspecific cDNA synthesis, hemi-nested PCR was carried out. The first PCR reaction was done with a spliced leader primer named ME33 (5'-AACACATCTATTCA-3') and the P2 primer. Position of the second PCR primer internal to P2 position named STOP (5'-TGCAGCAGAGG-GTGTACG-CC-3') because it contained the stop codon of the genes, were used. PCR products were ligated into pGEM-T vector (Promega). Positive clones were identified by hybridization with \textit{MUC.CA-2} probe in high stringency conditions (0.1x SSC, 65 °C). Thirty clones obtained with primers ME33-STOP and ME33-P2 were sequenced.

**Southern and Northern Blots—**Southern and Northern blots were as described elsewhere (20). Both were transferred to Hybond-N nylon membranes (Amersham Life Science, Inc.) and UV cross-linked. The hybridization solution was 0.5% SDS, 5× Denhardt’s solution, 100 μg/ml salmon sperm DNA, 3× SSC (0.45 mM NaCl, 0.045 mM sodium phosphate buffer (pH 7.2), 7% SDS, 1 mM EDTA, and 100 μg/ml tRNA at 65 °C. The filters were washed in 40 mM sodium phosphate buffer (pH 7.2) at 65 °C for 2 h. The screening yielded 48 positive clones of different strengths in hybridization signals. Out of these, 29 were fully or partially sequenced.

1. J. O. Prevato and L. Mendoza Prevato, personal communication.

2. The abbreviations used are: PCR, polymerase chain reaction; RT, reverse transcription; UTR, untranslated region; nuc, nucleotide(s); ORF, open reading frame; HV, hypervariable.
containing TcMUC genes. DNA of positive cosmids from a T. cruzi-ordered library (22) was digested with EcoRI, electrophoresed in a 0.8% agarose gel, and transferred to a nylon membrane. Hybridization was carried out using complete gene MUC.CA-2 (14) as the probe. Washing conditions were 0.1× SSC, 0.1% SDS, 65 °C. Eleven representative clones of the total analyzed are shown. kbp, kilobase pairs.

EcoRI site. In every lane, at least two bands large enough to contain a whole gene gave strong hybridization signals; in addition, several other weaker bands were observed in each clone. Consequently, we estimated a minimal number of 2 mucin genes per positive cosmid to obtain a minimal, and almost surely underestimated, gene number of 484 per haploid genome.

Similar hybridizations were made on the library filters using central regions of different cDNAs from mucin genes (described below). A probe derived from one of the genes whose central domain is made up of repetitive amino acid units detected 32 cosmids per haploid genome, indicating that an important proportion of the mucin genes has this kind of central region. A second group of mucin genes has central domains lacking repetitive units. Three central region probes (EMUCt-5, EMUCCe-1j3, and EMUCt-6; see below) that do not cross-react in hybridizations under stringent conditions gave 1–3 positive cosmids/haploid genome. Thus, we concluded that they likely detected single-copy members within the family.

The unusually high number of genes found and the diversity suggested by hybridization experiments raised the question of which of them were actually transcribed in a single parasite clone and what was the structure of the transcripts. To answer these questions, 59 cDNA clones were analyzed. They were identified in a cDNA library (29 clones) or obtained by RT-PCR (30 clones) as described under “Experimental Procedures.” The overall structure of the clones and that of the deduced proteins is described in the next sections.

Untranslated mRNA Regions Are Highly Conserved and Define Groups of Transcripts—The analysis of 36 5′-UTRs allowed us to group transcripts into two main subfamilies, which were named types A and B. This region comprises the first base after the spliced leader to the last base before the translation initiation codon. Both groups differ in length and share no significant sequence similarity but are very conserved within each group. The type A 5′-UTR shows three different lengths of 19, 34, and 29 nt due to point mutations that change the starting Met and to the use of different trans-splicing acceptor sites (Fig. 2). Using the same criteria, the type B 5′-UTR varies between 41 and 49 nt. The A group was much more frequent than the B group; 30 and 6 members of each kind were identified, respectively. In both groups, alternative trans-splicing acceptor sites were found (Fig. 2), but as the genomic sequence is not available, it is not possible to determine if an alternative AG acceptor site is being used or the major acceptor site is not conserved.

On the other hand, 3′-UTR showed much more similarity among all studied sequences (not shown). Sequence comparison of all available 3′-UTRs (29 clones) rendered a fairly homogeneous group. The length from the stop codon to the poly(A) start was about 350 nt. The percentage of homology ranged from 60.2 to 88% when the first 250 nt of these regions were aligned. The last portion, before the poly(A) attachment site, of all 3′-UTRs studied was always very rich in uridine.

Coding Region: Presence of a Hypervariable Mature N Terminus—The coding sequence of MUC transcripts (Fig. 3) in all cases showed a sequence previously postulated to be the signal peptide for endoplasmic reticulum import (14). This region is highly conserved, with amino acid identities from 80 to 100% among different variants. This peptide, whose size ranges from 23 to 28 amino acid residues, is very rich in Cys, with 6 of these residues in conserved positions. Cleavage site predictions based in current models (25) suggest the 6th of these Cys residue as a good candidate. It is therefore used in this analysis as the border between the signal peptide and the putative mature N terminus. Additionally, this Cys was conserved in all but one of 39 sequences analyzed.

The next few amino acids after the signal peptide would constitute the mature N terminus of the proteins once expressed in their final cellular destination and shown to be the most interesting region of the protein. Fig. 4 clearly illustrates the variability existing in this region that led us to name it the hypervariable (HV) region. For repeats-containing sequences, it was defined as the amino acids between the 6th Cys residue and the start of repeated elements. The lengths of these HV regions varied between 7 and 17 amino acids. In nonrepeated products, to facilitate comparisons this region was arbitrarily defined as the 15 amino acids following the signal peptide last residue. The sequence of HV regions was available in 34 of the isolated clones, from which 28 variants were determined; some of them being obviously related but gradually diverged to be totally different in some cases. In Fig. 4, all the variants available from CL-Brener are compared, arranged to show the gradual divergence. A common feature of all these sequences is that they are highly polar and hence hydrophilic. All of them have at least one and very often more than one acidic residue (both Asp and Glu). The presence of a positively charged residues (most frequently Lys) is also common in most of them as well as a clear abundance of uncharged polar amino acids (Thr, Ser, Asn, and Gln). The origin of this diversity is addressed in the next section.

After the HV region, there is a central domain having the target sites for O-glycosylation. Forty six out of 51 sequences isolated in which the central region was available differed in this central domain. The deduced amino acid sequence allowed classification of clones into the two main groups we reported earlier (14), those having central domains composed of tandemly repeated elements and the ones lacking repetitive central domains. These results clearly demonstrate the intrastain diversity for the central region of TcMUC genes. Twenty-seven clones coding for proteins with repeats representing 23 different variants for the central region were identified. The consensus repeat has the sequence Thr-Glu-Pro, but other repeats with the sequences Asn-Thr-Lys-Pro, Thr-Lys-Ala-Pro, or Thr-Gln-Ala-Pro were also identified. Codon usage for Thr indicated a high rate of synonymous mutation in this region. The deduced sequence from other 24 clones in which central regions were determined didn’t present repetitive units. From this last group, 23 variants can be identified having regions with scarce homology (less than 25% amino acid similarity).
Besides being rich in Thr and Pro residues, central domains lacking repetitive units are also rich in Ser residues. In fact, the ratio Thr/Ser is around two in genes lacking repeats, whereas Ser residues are hardly found in any of the central regions of genes with repetitive motifs. One of these nonrepetitive variants represented by three clones isolated independently by both strategies was totally new and delineated a third group within TcMUC genes. The deduced protein sequence lacked the degenerated repeats and Thr runs that all the others genes had (see Fig. 3 for a comparison of the three groups and their features). However, members of this third group conserved the same overall structure in terms of amino acid content (34% Ser + Thr, 10.6% Pro, and 27.6% Ala + Gly, for the protein excluding signal peptide and glycoprophosphatidylinositol anchor signal) and hydrophilicity profile.

Finally, all the clones except those in the third group mentioned before (see Fig. 3) coded for a conserved C terminus highly homologous to those previously described (14). Due to the high number of sequences analyzed, it can be further dissected in two regions (Fig. 3): a hydrophilic Thr-rich region that varied from 10 to 24 amino acids because of differences in the number of Thr residues and a hydrophobic region of 36 to 38 amino acids. The latter contains an amino acid stretch with the features of a glycoprophosphatidylinositol anchor signal (26).

Genetic Diversity Origin in Mucin Genes: Evidences of a High Mutation Rate in Localized Regions—Evidences of localized mutation came from comparison of two highly homologous mucin sequences. In a previous work we reported the genomic clone MUC.RA-1, isolated from the RA strain of the parasite (13). Comparison of the protein deduced from MUC.RA-1 with that of EMUCt-6 showed a high degree of identity in most of the central region. As can be seen in Fig. 5, an amino acid alignment clearly showed that differences begin at the HV region, extending toward the C terminus but gradually recovering an almost perfect identity within the central domain.

Alignment of the nucleotide sequences encoding the HV regions showed different degrees of homology. Examples are illustrated in Fig. 6. In some cases, minor changes were ob-
served, like conservative changes due to synonymous point mutations and small differences in the number of amino acids composing the HV region (Fig. 6, panel I). However, in many cases changes were more pronounced because of a higher number of nonsynonymous point mutations and/or insertions/deletions of nucleotides. This is especially interesting in genes having amino acid repeats that, besides differences in the number of repeats, were almost identical except for the region encoding the HV region (Fig. 6, panel II). In these genes, a localized accumulation of mutations was seen. More drastic changes include frameshifts due to insertion/deletion events (Fig. 6, panel III). These even lead to the premature appearance of stop codons in one of the MUC gene frames analyzed (Fig. 6, panel III).

In fact, four cDNA clones in the group of mucin genes lacking a repetitive central domain have stop codons in the supposed ORF and no other alternative ORF (not shown). They still conserve the nucleotide high homology in the regions corresponding to UTR and the 5' - 3' -ends, where even the reading frame was conserved. Mutations giving rise to premature stop codons in all four pseudogenes were localized in the hypervariable and central regions. The possibility that these clones represented transcription from pseudogenes will be addressed later (see “Discussion”).

In Table I we show nucleotide homologies and deduced amino acid identities, comparing the different domains of some nonrepetitive clones. In the central domain, they were more conserved at the DNA level (about 40%) than at the protein level (about 17%), whereas homology and identity values were much more similar in the other regions of the genes and proteins, respectively. Even clones with no ORF, like EMUCt-6, gave similar values. Sequence analysis showed that these differences were the result of accumulated point mutation, but also of small insertions and deletions all over the HV and

Figure 4: Deduced amino acid sequences of hypervariable regions from the CL-Brener clone of T. cruzi.

Sequences are separated in three groups according to the classification done in Fig. 3. Amino acid identities are indicated by colons and amino acid similarities by italics, and gaps introduced for best alignment are indicated by dashes. Asterisks indicate stop codons. The names of the clones with each sequence are listed at the right. Underlined names indicate identical clones. V.N.T.R., variable number of tandem repeats; N.R., nonrepetitive; C-TERM., C terminus.

Figure 5: Amino acid alignment of deduced proteins from clones MUC-RA-1 and EMUCt-6. Clone MUC-RA-1 (14) was isolated from strain RA of T. cruzi. Identities are shaded. A degenerate repeat is indicated by a box. The arrow indicates the border between the Thr-rich region and glycoprophosphatidylinositol anchor region. Alignment was done using clustal method (43) in the Lasergene package from DNAstar INC.
FIG. 6. Alignment between pairs of hypervariable regions. Nucleotide homologies are indicated by a bar. Amino acid identities are indicated by bold letters. Asterisks indicate stop codons. Gaps introduced for best alignment are indicated by dashes. Clones with repeats in the central domains are indicated by [R] and those lacking repetitive units by [NR].

TABLE I

Nucleotide homology and deduced amino acid identity comparisons between clones having nonrepetitive central regions

Values are indicated in percentages of nucleotide homology and, where corresponding, percentages of amino acid (aa) identities for different regions of the clones, as calculated after alignment of sequences using the Lasergene package of DNASTAR Inc.

| S' UTR | Signal peptide Hydrophobic | Thr-rich Glycosylphosphatidylinositol |
|--------|--------------------------|-------------------------------------|
| EMUC-6  nt 91.1 94.2 46.4 78.6 92.7 |
| EMUC-12 aa 91.5 20.9 71.4 85.7 |
| EMUC-6  nt 82 95.4 45.2 64.1 97.5 |
| EMUC-17 aa 100 21.6 58.5 84.6 |
| EMUC-12 nt 85.3 89.8 41.4 41.8 91.3 |
| EMUC-17 aa 91.3 10.1 59.3 84.6 |
| EMUC-6  nt 94.2 95.4 39.9 58.7 92 |
| EMUCe-1j3 aa no ORF in EMUCe-1j3 |

* Lower values in this region are mainly due to size differences.

FIG. 7. Northern blot analysis using individual central region probes. Panel A, ethidium bromide staining of ribosomal RNA bands from total T. cruzi epimastigotes RNA electrophoresed in an agarose gel. Panel B, Northern blot hybridizations corresponding to the Panel A gel. Probes correspond to clones EMUCt-5 (central-t5), EMUCt-6 (central-t6), and EMUCe-1j3 (central-1j3) as described under “Experimental Procedures.” Arrowheads indicating the position of the bands correspond to 1100 bases (1), 1000 bases (2), 850 bases (3), 750 bases (4) as calculated using RNA size markers (Life Technologies).

DISCUSSION

The T. cruzi mucin gene family and their transcription units were shown to be large and complex in a single parasite clone. A minimum of 484 genes are present/haploid genome. Even though this is a fairly minimal estimate, this figure indicates that about 1% of the parasite genome is occupied by mucin genes, assuming 1 kilobase pair/coding unit and 55 megabase pairs/haploid genome (23). Analysis of the UTRs in a number of different transcription units showed that members of the family are likely to have a common origin. In particular, the 3' UTRs are very similar in all the cDNAs analyzed. UTRs are, in fact, much more conserved than some of the mucin coding regions. Whether this is due to sequence restrictions responding to requirements for mRNA processing, stability, or regulation of stage-specific expression remains to be tested (27). Conversely, the mucin central coding region might evolve rapidly, since it is only restricted to conservation of the target sites for O-glycosylation.

Two groups of genes were identified according to their central coding regions, those having and those lacking repetitive units. Since relics of repetitive units are detectable in the latter group, it is possible that the ancestral gene was the one having Thr- and Pro-rich repetitive units. Once the organization into repeats is lost, no sequence homogenization by unequal recombination or gene conversion can occur, leading to genes having...
Diversity of T. cruzi Mucin Transcripts

unique central regions. This is in fact what was observed. Genes bearing repetitive units form to a large subfamily according to the genomic library hybridization and the percentage identified in the 59 cDNAs analyzed. Given the high synonymous mutations observed in the Thr repeat units, it is possible that this particular group is under strong selection pressure. On the other hand, genes lacking the repeats have unique central domains shared by few other members, even though this group is probably conformed by hundreds of genes. Genes coding for repeats have, almost exclusively, Thr residues as possible target for O-glycosylation. Synthetic peptides made according to the consensus sequence of these repetitive units have been recently shown to be excellent substrates for an UDP-N-acetylglucosamine:peptide:N-acetylglycosaminyltransferase.\(^1\) Thr seems to be the main target for O-glycosylation in T. cruzi \(^2\).\(^3\) However, the mutations accumulated in the genes bearing nonrepetitive central domains generated in addition a substantial number of codons for Ser residues as possible targets for O-glycosylation, the hydroxyamino acid content for the region corresponding to the mature protein being more than 30% in all the clones analyzed. The ratio of Thr/Ser codons in the same region is around 2 in nonrepeated clones, far from the value around 7 in repeats containing clones. These results would suggest the existence of more than one N-acetylglycosaminyltransferase in T. cruzi, as shown in other organisms for enzymes initiating O-glycosylation \((28)\).

Sequence analysis suggested the hypervariable region to be a hot spot undergoing substantial mutation and hence generating large diversity in the corresponding protein region. We found that at least four clones coming from the cDNA library were mature transcripts, but the expected coding frame was interrupted by stop codons. Despite the presence of stop codons, these clones are very similar in sequence to one or more of the mucin genes identified. Interestingly, the mutations generating the stop codons are located in the hypervariable and/or central coding region (see below), whereas the UTRs and 5′- and 3′-ends are practically identical to those present in functional genes. Since cDNA clones were analyzed in this work, it was unexpected that clones lacking an ORF were isolated, suggesting transcription from pseudogenes. This finding might be the consequence of the rapid evolution of sequences within coding regions, resulting in the appearance of pseudogenes \((32)\). Although further work on the corresponding genomic loci and protein product is necessary to confirm the existence of these pseudogenes, one likely explanation for the isolation of cDNA clones without ORF would be that mutations in the coding region giving rise to pseudogenes were very recent events, still allowing a mature mRNA to be synthesized but not translated into a functional protein. Altogether, these findings would suggest a genetically active state of these loci with a high mutation rate localized in the hypervariable and/or central regions.

There are few examples in protozoan parasites of large and variable gene families coding for proteins or glycoproteins with a similar or identical function. Among them are the ones encoding the variable surface glycoprotein in African trypanosomes \((\text{see review in Ref. 29)}\) and the ones coding for the var genes in Plasmodium falciparum \((30, 31)\). In both cases, the large number of genes is directly related to a requirement for parasite survival in the host, and sequence variability is the naturally selected character. The N-terminal region of these proteins, the ones exposed to the host immune system, largely differs among members of these families. Changing from one antigen to another is the mechanism used by these parasites to evade the effective antibody response against themselves in the case of African trypanosomes or infected erythrocyte clearance in Plasmodium infections \((\text{reviewed in Ref. 32)}\).

In the case of T. cruzi mucins, in addition to the O-linked sugars attached to the central domain, there is a second region likely to be exposed to the host. This region is located in the N termini of the mature glycoproteins, and it was found to be hypervariable and highly hydrophilic. Its peptide sequence might be directly exposed and not covered by carbohydrates since it generally lacks a consensus target sequence, as inferred from computational predictions \((33)\). This region is indeed hypervariable, as demonstrated from the sequence of a number of cDNAs and, in an extreme situation, by the fact that two cDNAs that are almost identical vary greatly in the hypervariable region \((\text{Fig. 4)}\). It is possible that this region is being selected for variation. The diversity generated might allow the parasite to find a receptor with which to interact in the host mammalian cell surface. T. cruzi is known to have the ability to invade almost any cell in its hosts \((34)\) as well as to have a wide host range \((35)\). Another alternative is that this hypervariability is required to generate a large and diverse set of epitopes exposed to the immune system. These sometimes related but different epitopes would in term preclude the maturation of the immune response against a restricted set of sequences \((36)\). Finally, it cannot be excluded that one or few mucin classes were expressed at a time, generating a mechanism of antigenic variation to evade the host response. In any of these cases, our model proposes that the glycosylated central domain in mucins would constitute a stalk extended on the parasite surface, exposing the hypervariable region on top. Extended mucin-like stalks have been found in a number of membrane-bound molecules \((37–39)\) and can be predicted because the high Pro content prevents an organized secondary structure, and the O-glycosylations flattens the molecule.

Another reason for a large number of mucin glycoproteins in T. cruzi is the one previously proposed by several groups, their involvement in the invasion of mammalian cells through their sugar components. Antibodies to the sialylated Ssp3 epitope in the parasite mucins are known to block parasite invasion \((8, 11)\). However, recent results cast some doubts on the requirement of sialic acid in mucins for the invasion of host cells. In fact, it was suggested that the absence of sialic acid favors parasite invasion into mammalian cells \((40)\). Thus, it is still possible that sialic acid interferes in the host parasite interaction due to its negative charge and that other parts of the mucin molecule are required to bind to mammalian host cell surface structures.

Although a large and diverse set of cDNAs were identified, all of them might not be translated into glycoproteins. In fact, four of the mRNAs identified are likely to be transcripts from pseudogenes. Trypanosomatids are known to regulate expression at different levels from the initiation of transcription \((41)\). We have found preliminary evidence that the relative abundance of different transcripts might vary in the epimastigote stage of the parasite \((\text{Fig. 7)}\). Developmental regulation of the TcMUC gene family has been observed among stages of T. cruzi \((14, 42)\). Biochemical studies showed that there is an heterogeneous group of 35–50-kDa mucins that is essentially expressed in the forms of the parasite present in the insect vector \((3, 10)\). The second group of mucins is a large variety of molecules ranging from 80 to 200 kDa expressed in the parasite stage that is infective to mammalian host cells \((6)\). Thus, there is strong evidence indicating that the large family of mucins are developmentally regulated in their expression. The question of how many different members are simultaneously expressed as glycoprotein in each group is not known. Our evidences suggest that members of the TcMUC gene family code for the core protein of both mucin groups; the family of 35–50 kDa \((13)\) and
the one of 80–200 kDa. We show here that all genes might be simultaneously transcribed in both epimastigote and trypano-
mastigote stages. Further work is required to know whether all
transcripts are simultaneously translated, which ones are ex-
pressed in the different parasite stages, and their functions.
Identification of the hypervariable regions together with
transcripts are simultaneously translated, which ones are ex-
pressed in the different parasite stages, and their functions.
Identification of the hypervariable regions together with
transcripts are simultaneously translated, which ones are ex-
pressed in the different parasite stages, and their functions.
Identification of the hypervariable regions together with
transcripts are simultaneously translated, which ones are ex-
pressed in the different parasite stages, and their functions.
Identification of the hypervariable regions together with
transcripts are simultaneously translated, which ones are ex-
pressed in the different parasite stages, and their functions.

Acknowledgments—We thank Dr. J. Hoheisel (Heidelberg, Germany)
for the high density arrays filters and the T. cruzi CL-Brener cosmid
library and Dr. E. Rondinelli and Dr. T. P. Urmenyi (Rio de Janeiro,
Brazil) for providing the epimastigote cDNA library to Åslund. We
also wish to thank Dr. J. J. Cazzulo for critical reading the manuscript
and Liliana Sferco for parasites cultures.

REFERENCES
1. Gum, J. R., Jr. (1992) Am. J. Respir. Cell Mol. Biol. 7, 557–564
2. Jentoft, N. (1990) Trends Biochem. Sci. 15, 291–294
3. Previato, J. O., Jones, C., Gonçalves, L. P. B., Wait, R., Travassos, L. R., and
Mendoza-Previato, L. (1994) Insect Immun. 15, 921–925
4. Schenkman, M., de Almeida, I. C., Pereira, M. E., Ferguson, M. A. J., Travassos,
L. R., and Gazzinelli, R. T. (1997) J. Mol. Biol. 267, 2205–2218
5. Yoshida, N., Mortara, R., Arajuth, M. F., Gonzales, J. C., and Russo, M. (1989)
Insect Immun. 37, 2446–2449

3 G. D. Pollevick, J. M. Di Noia, C. Lima, R. Lederkremer, D. O. Sánchez, and A. C. C. Frasch, unpublished results.