Variant Specific Glycoprotein of *Trypanosoma brucei* Consists of Two Domains each Having an Independently Conserved Pattern of Cysteine Residues

Mark Carrington, Nancy Miller, Michael Blum, Isabel Roditi, Don Wiley and Mervyn Turner
Variant Specific Glycoprotein of *Trypanosoma brucei* Consists of Two Domains each Having an Independently Conserved Pattern of Cysteine Residues

Mark Carrington¹†, Nancy Miller¹, Michael Blum², Isabel Roditi³
Don Wiley² and Mervyn Turner⁴

¹Department of Biochemistry, Tennis Court Road
Cambridge, CB2 1QW, U.K.
²Department of Biochemistry & Molecular Biology
Harvard University, 7 Divinity Avenue
Cambridge, MA 01238, U.S.A.
³Institut für Allgemeine Mikrobiologie, Universität Bern
Baltzerstr. 4, CH 3012 Bern, Switzerland
⁴Merck, Sharp & Dohme Research Laboratories
PO Box 2000, Rahway, NJ 07065, U.S.A.

(Received 27 November 1990; accepted 19 June 1991)

The complete amino acid sequences for nine variant specific glycoproteins (VSGs) of *Trypanosoma brucei* are presented. These have more than doubled the size of the VSG sequence data base and have enabled a new and more rigorous comparison to be made between amino acid sequences of different VSGs. Each VSG can be defined as a combination of an N-terminal domain type and a C-terminal domain type, based on the distribution of cysteine residues within the molecule. This identifies three N-terminal domain types and at least four C-terminal domain types. Different combinations of N and C-terminal domains can be formed; for example, in the sequences presented here, two different N-terminal domains are found in association with each of three different C-terminal domains. The biological context of the domain structure of VSGs is discussed.

*Keywords:* *Trypanosoma brucei*; VSG, sequence comparison; domains; cysteine conservation

1. Introduction

A population of the protozoan parasite *Trypanosoma brucei* can persist in the bloodstream of the mammalian host through the process of antigenic variation. The surface of the bloodstream-form of the parasite is covered by a 12 to 15 nm thick coat (Vlekken, 1969) composed of a single polypeptide species, the variant specific glycoprotein (VSG; Cross, 1975). Antigenic variation occurs when an individual trypanosome stops expressing one VSG gene and starts expressing another which encodes an antigenically distinct VSG (for reviews, see Page & Steinert, 1988; Horst, 1986) thus delaying clearance by the host immune system. Direct estimates have indicated that there are approximately 1000 VSG genes in the trypanosome genome (Van der Plouw et al., 1982) and although no estimate of the number of functional, antigenically distinct, VSGs that can be generated from these is available, infection of a rabbit with a single cell from the related species *Trypanosoma equiperdum* led to the appearance of more than 100 serologically distinct variants (Caperon et al., 1977).

Biochemical and sequence analysis of VSGs has shown that the mature polypeptides contain between 400 and 500 residues and have at least one N-linked oligosaccharide, although the position of this oligosaccharide varies (Cross, 1984; Turner, 1988). VSGs are found as dimers in solution (Aufré & Turner, 1981) and are believed to exist as dimers or higher oligomers at the cell surface (Brickner & Patton, 1982). Attachment to the cell surface is
through a glycosyl-phosphatidylinositol anchor linked to the C-terminal amino acid of the mature polypeptide (Cardoso de Almeida & Turner, 1983; Fergaon et al., 1988). An N-terminal domain can be physically separated through the action of exogenous proteases (Johnson & Cross, 1979). This domain contains two-thirds to three-quarters of the amino acid sequence and also exists as a dimer in solution. The N-terminal domain has now been crystallised for several VSGs (Metcalfe et al., 1988) and the crystal structure has been solved at high resolution for two of them (Freyman et al., 1984, 1990; Metcalfe et al., 1987; M. Blum et al., unpublished results). The C-terminal domain contains the balance of the sequence. At present, the structure for neither an intact VSG nor an isolated C-terminal domain has been reported.

The complete sequence of seven T. brucei VSGs have been published, derived either directly from the polypeptide (MITat 1-4; Allen et al., 1982); and/or from the sequence of cDNA (MITat 1-4, Boothroyd et al., 1982; ANTat 1-1, Michiels et al., 1982, Pays et al., 1983; ANTat 1-10, Pays et al., 1985; MVAT 4, MfAT 7, Leonardo et al., 1984; ILTat 1-3, Rice-Ficht et al., 1981) or a combination of cDNA and genome clones (ILTat 1-1, Rice-Ficht et al., 1982). It should be noted that (1) ANTat 1-1 and ANTat 1-10 have >75% amino acid sequence identity and are encoded by members of the same gene family despite not cross-reacting antigenically when expressed on living trypanosomes and (2) a partial polypeptide sequence, derived from cDNA sequence, is available for ANTat 1-8 (Matthyssons et al., 1981), this has 104/115 identities with MITat 1-4 and so both are probably members of the same VSG gene family.

Comparison of these sequences has shown that the sequence of the N-terminal domain is extremely diverse except for the relative locations of cysteine residues which are conserved (Cross, 1984). The C-terminal domain exhibits a degree of sequence conservation (Rice-Ficht et al., 1981, Holder & Cross, 1981, Engelund et al., 1988). This homology is found both near the C terminus of the mature polypeptide, and in a hydrophobic C-terminal extension present in a VSG precursor, detected by comparison of cDNA and polypeptide sequence data (Holder & Cross, 1981), that presumably acts transiently as a membrane anchor before glycosylphosphatidylinositol addition. As in the N-terminal domain the conservation of cysteine residues is particularly strong in the C-terminal domain of the mature polypeptide.

In this paper a further nine VSG sequences are presented. These additional data have enabled a more coherent analysis of conservation in the primary sequences of VSGs. Each VSG can be described as a combination of an N-terminal domain type and a C-terminal domain type based on the distribution of cysteine residues. The N-terminal domain types and C-terminal domain types are interchangeable so that, in different VSGs, one N-terminal domain type can be found in combination with most C-terminal domain types and vice versa.

The domain structures for nine members of the ILTat 1 serodeme (Miller & Turner, 1981) are therefore known, and the relationship between domain structure and switching can now be described for this serodeme.

2. Materials and Methods

(a) Trypanosomes used

All trypanosomes used originated from the MITat 1 (Cross & Manning, 1973) or ILTat 1 (Miller & Turner, 1981) serodemes.

(b) Oligodeoxynucleotides

Oligonucleotides were synthesized using a Biosearch cyclone DNA synthesizer. Oligonucleotide 39 (5'TGGTTAAACTATATCA3') was designed to base-pair with a conserved sequence nucleotide sequence found in the 3' untranslated region of VSG mRNA (Majumder et al., 1981). Oligonucleotide 41 (5'GGTTCTGATATATTTT3') was designed to base-pair with the reverse complement of the 3' part of the mini-exon found at the 3' end of all T. brucei mRNAs (Cornelison et al., 1986).

(c) Preparation and sequencing of VSG cDNAs

Polyadenylated RNA was purified from bloodstream trypanosomes (Roditi et al., 1987). One of two procedures was then followed: either a cDNA library was constructed in λNM1149 (Roditi et al., 1987) or double stranded VSG cDNA was prepared using primer extension with oligonucleotide 39 followed by the polymerase chain reaction (PCR) using oligos 39 and 41 (Kawasaki, 1990). Clones from cDNA libraries containing cDNA derived from VSG mRNA were identified using 32P-oligonucleotide 39 and, if necessary, further clones isolated by screening with purified cDNA from initial isolates. PCR products were cloned into a plasmid vector. All cDNAs were sequenced by generating random subclones by somatic (Rae et al., 1984), which were cloned into M13 mp18 and sequenced by the dyeex dye method (Sanger et al., 1977).

3. Results

(a) Synthesis and sequencing of cDNAs encoding VSGs

The cDNA sequences are contained in the LMBL database, accession numbers are shown by each amino acid sequence in Figure 1. With the exception of MITat 1-2, these cDNA sequences contain part or all of the T. brucei mini-exon at their 3' ends. In the case of the MITat 1-2 the first 16 bases of the sequence are taken from previously published data (Boothroyd et al., 1981). From results using primer extension there are approximately 45 bases 5' of the sequence presented before the mini-exon (data not shown). At or near the 3' end of each cDNA sequence the conserved region found near the 3' end of VSG mRNAs (Majumder et al., 1981). Taken together these two sequence blocks indicate that the data presented represents complete VSGs. The origin of the cDNAs sequenced were as follows: MITat 1-2, ILTat 1-24 and ILTat 1-25, cDNAs
MITat 1.1 (MIAG 060) EMBL X56761
MATGRAKNTK WARNLSTAGL IVVTLPATT MAAPRTGTLKA TAATAKLFCLT TELSKVSGEM 60
INKQGEVISN IQKIAAYEY VSYYLAKPE TQALQGSTLL RGYFARKTNG GLEYSKMGL 120
ATQRSARAA AYLKSIDMF LNNESSLQG SENKCLVTTN ADTAATRRET KLDQECALS 180
MPETKPSAAT RTELCTQYP ILQHGSGGGA NTQFQPTSTG TCKLSSHGHT NGYPPSSALD 240
TTAKVLGAYM TIPNTQVEAT LANMQAMNGN HKATAPAWHE AWEARNREAK AKDILAYNET 300
GNLDTQPTLK ALVKTLLLPK DNTENHAEAT KLEAFGGGLA ADKTTYLDM VDAEIIPAGI 360
AGRTTEAPILK KINHTVELGD ILSHYEMIAA QNVTLKKNL DAVSKQQTe SAEKKEKICN 420
AAKDNDQACE NLKEKGCYFN TESNKCELKK DVEKEKLEKES KETEGKDEKA NTIGNSFEL 480
HKEPLLLAFELF

MITat 1.2 (MIAG 221) EMBL X56762
MPSENEARLF LAVLIVLAQVL PbLVDSAAEB EFKQAFWQQPL CQVEEELDDQ PKGALFTLQA 60
AASKIQKMD AALRASYAE INHGTNRAKA AVIVANHWM KADSGLALEK QTILSSQEVT 120
TATASYLKG ILEYNLLLQ TKEGTSGCCM MDTSGNTVT KAGGTIGGV PKLQISIPQI 180
KRPAATYLGG AGYVGLTROA DAAANNHPHNDN AECLASGSN TNGLKRSGQL SAAVTMAAGY 240
VTYANSQTAV TVQALDALQS AGSAAIFEQW DAWKAKKALT GASTABFRNE TAGIAKGTGV 300
TKLVEEALLK KKDSEASEIQ TELKKFESGHH ENEQWTAIEK LISESQPVAQI LVGDNQTPKI 360
GELEGPAKLTL TILAYRME#T AGKEFLQEBK HKPAESQQQA ABTEGSCNKK DQNECKSPCK 420
WHNDAEKNNKC TLDKEZAKKV ADEYAKDKQT GHDNTIGGSSN SEVISRTFLN LAVLLEF

MITat 1.5 (MIAG 118) EMBL X56763
MIHNSKVATV VLALISSWSPA GTGNNHGKRL QRAAQICKMS KELKATAMRA ANDAHLKITE 60
ILELENVFAA MIPNATKGTK ADGCTDNAY FLEAANNTAE TVSKRIATLAE SATKAAGGAAG 120
RAAGVLDFFI AALAAQAGQAT GLCIQSOGST GAATHTELAD CFNGLDKRPN MLSIRDQKVS 180
AAATGSTLTL TLAKAMASSG TDTPFHDQQQ SKGCQMLKGT SDGIMIQQAL TGTFWAQQL 240
LRFUGLGA ENGASTVYGA HYTAASGNOV HWAASDEKIP VIAEATALVVS NNYLMADSIG 300
TRAKDAIEKV KCCMKATINKE IKREHIFLNV SHLNRQOLKA VTLKDKALNK QDAKAEAKQQ 360
PNCDDKKQIE CGDTPGCNNW KAEGKCEAKD GGEQQQATGQ EKDAKNOCTV QHGTNKEACE 420
KENTFGQAV CGFVRKGKDGTE TDEPDKEKCR NGSPITSKOF APSVVSAAEFL ALLEF

Figure 1. The amino acid sequences of VSGs from the MITAT 1 and HLTA 1 serozone. Each sequence is identified by the encoded VSG and the MIAG clone number of the trypanosome population and then the accession number of the cDNA sequence in the EMBL database. The N-terminal signal sequence and the C-terminal hydrophobic extension absent in the mature polypeptide are underlined.
MITat 1.6 (MIAG 151) EMBL X56764
NAVIRALAAV AISLYVLLFR KGATDKGAI KFETWEPLCL LTQDFGNLYN RAHKLNLID 60
TYVTAAQADQ LRLQILLSRA SSKIEAAAA AATAIAADL AGKAKHAVSC KLAATTITAT 120
TGYLHGRJA FLEVMTAHRG TTNKYGCLSK SRSDNSNSIT DSVANLKDCC KLTVOQISP 180
NKATEQITKA GFTKLATGGL LTSSDLGGSG QAHCNLSST ASEVNVNGL DEPVYAGGY 240
LRKRDHLLTD GNDNLIATID SASAPATRAK TDPLYQWIRA FKNLDECEST FTGSYSPSP 300
ETLPKAADETK TAIKNVYVVK EKYDWQATDK EDDYNLKDKI FRDCKDYPFQ KLVADNKLKD 360
LLKDATQITNE IKKLADITDR SELNKQLYY TRQEKQTITK ELKEAQERAT QANQNDAAAK 420
AARDSCNKLV GEGSKNADKX CSYETETGDT KKKCPNATRA EKSGAPVQTA QTVGTSATTP 480
EKCKGKDACT CGTTCQGCKWE GETCKDSLLV VTKEFALTYY SAAFVALLE

ILTat 1.21 (MIAG 201) EMBL X56766
MLBELLPLST TALALLAGCH AQVGDAFPAP AVLCAANWDA TAQKQKPWE DRELPEINDI 60
YNNNMSTASE EQWQTFQGQA EQQTSWQFAQ AAANGKYKGID WKNWWRKWRQ QRPQKTDAGG 120
AWQTKHHRPE WATFRPQVRP VILAIASAEAT ELSRKLEPPR TDAQDKLIAE INSKLASARC 180
SGELKAAAGN IGCTGPEGTP DKTTCTTAK AGGSICHMDL CICSVAEATD KCSSTGVGDA 240
VPNSGKLRNS NGFQHIVARQ PKPGESGTLP QAIDLALAML ATALGTOQPG SNNMLGKSK 300
GGCTATNSGA CVDYHERFSQ QQQITGIPW VALLQQARAL YGTIPDOKLA AQTARQIVYMN 360
LAGQAKREYR RPAGSLEKDPA GVIQEOFATNR RHRGADEYNQ CTSNNATADE CPETRCEYDS 420
EKNECRPKKG TETTATGPGE RITPADDGAK NTVSDSLPL IKSZELALAFELL F

ILTat 1.22 (MIAG 202) EMBL X56765.
MDTAQVFAPF YMATVMAATG KKNQAOVSD PCSEIHFDDEQ LANYFENAVS AATTQLDENQ 60
NPRRSHWFLQ YLQMDHQQSK GAAALAYAS TINIRTAANV KAAGSELLTA ASLLKQRAN 120
VSAAVFQOQSF GVIKLDTPDI DNGAKSITHA DAGCNRYAAS KTVPTQCRCTP PQQADITTA 180
ADMQPFKLLID LQNLITEAYTT TITIAASAYS KGTFTAQGTGTV VTYGNCOSTG GSASAQGDT 240
HALGIVKHTI OVTAKVETK LTQPSSNCMP DEGTTAELTP IZKLARACIL ARKSLAKPK 300
ALSRLQYSDL QTTDFKRIA AIFLRSNQKQ LDPEKSDQBI NEKILKETGF PMEEHFKSVY 360
EALDNHKWFEM KEKSKIGGT VNALANGYDA GLATAYYASK RQCSCQGAAAA DTPTVSSDE 420
KCKGKTQDBC RTADECMDR GECNAKAVKT AEPDSKNTTT GNSFNAIKTS TLLAVLLF

Fig. 1.
VSG sequence comparison

ILTat 1.23 (MIAG 206) EMBL X56768
MKKNINAVL LIIILSTRNDQ NAADAAGDENF LFLDLCLNLLE LGKRAVSKILA PTNLGRLAYS 60
EKIQKINMSLS DDAWAKAFFP KKKQKQENSNN ADGAPRSQEK THARNHARWQ AATDLAGDEG 120
DRKPTLRSLG LAVTVQVKIQYL SLALQPLABRAAIALEQLKTLHSGSAGLTSTNIRQEQIT 180
ALYGTGATAP EKTLLOLKK KKNVGSTRKD ICQGDNTAAK ADTVLALYLC ICGAHATDSG 240
GAIIKVCSTQ QPANNKADAV SDAHTHAAL AGQCQCSDDT NDKAAEIDS AILEFTSLKL 300
AAANQKPYFGK YSATGCTGSD AEIGICVMFTK TAKGERGKAVQ QIPWLTLHNAAMEIRKQQ 360
VNGKIDSINQ ELIQAIQTAAY ALPKQOLEMYK RLPQTTKEAK PKQQLTEMQAGECNTHKSN 420
TCPKNNCKWE EKDKGDQKCV ADDSKVTQGQ NAPAGAGDTD GTITTTTPCASONHTDTSCE 480
EMKGTTPPCG CNKSTRGEGES DQDKEMCENG FPFLAKKEFLAVSVSAAFTAL 4

ILTat 1.24 (MIAG 209) EMBL X56767
MVYRNLOLS VLVKVLVLVLI VEATHFGVYK ELWQPECELT AERLRTAGVA KMKVENSLNS 60
FKTLELTQMK LIIITFSAKFPE SKEAALTDR LAALNTDLRA LKDNIANGID RAVRATAYAS 120
EAAGALFSGI QTLHADTDGT TYCLSASQGQ NGNAMASQ GCKPLALPHEL LTEDSYNDV 180
TSDFKGPQKIS PITNAOOGQS GEGCGLOFAA SGATQTGNV QPSGGRINL GLGATVASAA 240
QQPQRPIDSD FSQXTNAQQAQD TLYGKARASI TELIQLAQGPKPQGTEVTKM ILIAQKTAAL 300
DSIKFQLAAS TGGKSDYKE DNLKTEYFG KTESNIRALW NVKKEKVEKG ADRPDPSKES 360
KISINDTEEQ LQRVLDYAYV ATMLKLAKQAAEDIAKLETF ADLQKGKSPAA ECNKITEEPK 420
CSEEKICSWH KEVKAGEKNC QFNSTKASKS GYVTQQTQTA GATAEAECK CQGKSDCKSP 480
DOCAEQQGQK DSSILANKOF AI8VASADEV ALLF

ILTat 1.25 (MIAG 211) EMBL X56769
MOSQCQVEF ILLAINTU DAAAPTTAVNA RCEFGLLCRTLV RAEDNLEDGR SAGQAAKEVV 60
ALAAKIELIL ALNKHIEIL AAEEPAAFKR SERSDDTEAC KASKATTVSQQ AAQTYKTIRP 120
DEKIAIAFLLA ETTGDIIRSTF NVTLLIQTAA AASHARYFGE NTESRPALDK IKKAALGSP 180
AKGDNIPSCD QGRSAACGN TDQNAANSAA KRATAAALC CGGDNNTNQG DACFFQTTAD 240
INYAKKQGEV ERAWTEIRQ CKAGAAANIK TAAQIKAAA EALAIHQQRG EKKAVALLLG 300
AAQINNGAVD CDGSHANGKG SCVILSTSAKKYKvetPDWLN NAEAAIAADL EQEQYIEQDM 360
RKAARQILAL NSS1TTLIAA AVEPTKQPAA AAAAAPEKKS NFQKDCNINTN KRRDCEGKGD 420
CWSSTEATE GAFCKKPDGE GQVTSAAGAGD AGASDTEARK CSDKKKEEGEC KPNCKWDOQ 480
ECKDSSILANKOFALLSVAEDEVALLF

Fig. 1.
isolated from libraries; ILTat 1-21 and ILTat 1-22, partial sequence from cDNAs isolated from libraries and complete sequence from PCR generated cDNAs; MITat 1-1, MITat 1-5, MITat 1-6 and ILTat 1-23, PCR generated cDNAs. The ambiguity at residue 374 of ILTat 1-25 VSG is due to a base difference in the two library cDNA clones sequenced.

(b) Primary structure of the VSG polypeptides

The amino acid sequences of VSGs MITat 1-1, MITat 1-2, MITat 1-5, MITat 1-6, ILTat 1-21, ILTat 1-22, ILTat 1-23, ILTat 1-24 and ILTat 1-25 are presented in Figure 1. The locations of cleavage of the N-terminal signal sequence were determined for each VSG by comparison of the cDNA-derived sequence with that determined by direct N-terminal sequencing of purified VSG polypeptide (ILTat 1-21, Gurrnett, 1988, MITat 1-1 and MITat 1-5, P. Voorheis, personal communication; the remainder, Metcalf et al., 1988). The N-terminal signal sequences of all the VSGs are typical of those found on eukaryotic polypeptides co-translationally secreted into the endoplasmic reticulum and, as previously reported (Cross, 1984), do not show any great homology to one another. The C termini of the mature VSG polypeptides from the MITat 1 serodeme were located by comparison of the cDNA-derived sequence with the directly determined C-terminal sequence (Holker & Cross, 1981). The designation of the mature C terminus of the ILTat 1 VSGs was based on sequence homology with members of the MITat 1 serodeme.

(c) Sequence homology between VSGs, conservation of cysteine residues in the N-terminal domain

The amino acid sequences of the nine VSG N-terminal domains (nominally taken to be residues 1 to 350 of the mature VSG polypeptide) were compared with the University of Wisconsin Genetics Computing Group GAP programme which estimates percentage identity after the introduction of gaps to maximize homology (Dovereux et al., 1984). Of the comparisons 34/36 gave values below 25% identity, MITat 1-1 and MITat 1-2 have 27.8% identity and MITat 1-2 and ILTat 1-24 have 246% identity. Since random polypeptides aligned in this manner give values of 10 to 20% identity (Doolittle, 1986) and the nine VSGs have similar amino acid compositions these values for percentage identity indicate that there is probably little homology between the N-terminal domains of the VSGs presented here and so they represent nine new examples of functional VSGs.

The location of cysteine residues in the N-terminal domain divides the VSGs into three groups based on the relative positions of these residues (Fig 2(a)). All the VSGs have a conserved cysteine residue between positions 13 and 17 from the mature N terminus, however, other than this cysteine each of the three groups appears to have a distinct pattern of conserved cysteine residues. The majority of VSGs, including all the members of the MITat 1 serodeme sequenced thus far, are type A and are characterized by the four conserved cysteine residues. Of the eleven type A VSGs, six have just these four cysteine residues in their N-terminal domains (MITat 1-1, MITat 1-2, ILTat 1-3, ILTat 1-24, ANATat 1-1 and ANATat 1-10). Three of the others have an extra pair of cysteine residues the position of which is variable (MITat 1-6, MITat 1-6 and ILTat 1-22), MITat 1-4 has one and MVAT 4 has three additional cysteine residues.

There are three N-terminal domain type-B T. brucei VSGs (ILTat 1-21, ILTat 1-23 and ILTat 1-25): these VSGs have eight conserved cysteine residues. ILTat 1-21 and ILTat 1-25 each have an additional pair of cysteine residues, the position of which is not conserved. Also shown in Figure 2 are two VSGs (YNat 1-1 and YNat 1-3) from Trypanosoma congoense (Strickler et al., 1987) which are truncated when compared with VSGs from T. brucei, as they have no C-terminal domain. In addition to the eight conserved cysteine residues, an additional pair occurs in both ILTat 1-21, ILTat 1-26 and YNat 1-1; and a single additional cysteine residue occurs in YNat 1-3. None of these residues occurs in a position that is conserved between different VSGs.

There are two T. brucei VSGs that have been classed as N-terminal domain type C (ILTat 1-4 and MVAT 7) as has VSG 20 from T. equiperdum (Thon et al., 1990). These have six conserved cysteine residues; ILTat 1-4 has an additional two cysteine residues not found in MVAT 7.

The potential N-linked glycosylation sites (Asn-X-Thr/Ser) are shown in Figure 2. Generally there appears to be little conservation in the position of these sites in the primary structure of the N-terminal domains.

(d) Sequence homology between VSGs, conservation of cysteine residues and glycosylation sites in the C-terminal domain

VSGs have previously been divided into two groups based on the number of cysteine residues in the C-terminal domain; group 1 VSGs have eight cysteine residues and group 2 have four or five (Rice-Ficht et al., 1981, Turner, 1988). Figure 2(b) represents the primary structure of the VSGs aligned on their mature C terminus showing the location of cysteine residues and potential N-linked glycosylation sites. With the additional data available the VSGs have been grouped into four proposed C-terminal types. These groupings are based on: (1) the number of and location, relative to the C terminus, of cysteine residues; (2) the identity of the C-terminal residue; (3) the sequence of the C-terminal hydrophobic extension and (4) the position of potential glycosylation sites. Figure 3 shows the amino acid sequences of the C-terminal and hydrophobic extensions of these VSGs.
Figure 2. Representation of the primary structures of VSGs for which the complete amino acid sequence is known. The mature polypeptide is shown, the location of cysteine residues are marked by vertical bars and potential N-linked glycosylation sites by the filled circles. Disulphide linkages are shown on MTat 1-4, the only VSG for which these have been determined. The position of the domain boundary is approximate and is based on the size of fragments produced by mild proteolytic digestion (Johnson & Cross, 1979; Metalf et al., 1988) and on the results of the crystal structure solved for MTat 1-2 (Freyman et al., 1980) and MTat 1-24 (Blum et al., 1990). (a) The sequences aligned at their N termini to show grouping into N-terminal domain types. The arrows above the 1st member of each group show the cysteine residues conserved within that group. (b) The sequences aligned at their C termini to show grouping into C-terminal domain types.

Eight VSGs have type 1 C-terminal domains containing eight cysteine residues in two groups of four. Each of these groups of four forms two internal disulphide bonds in MTat 1-4 (Allen & Garnett, 1983) and it is probable that the other members of the group form the same disulphide linkages. All the VSGs in this group have aspartic acid as the C-terminal residue of the mature polypeptide and have a hydrophobic extension of 23 residues that starts serine-serine. Type-1 C-
C-terminal domain type 1

** * * * *** * * * * * * * * * * * *

MITat 1.4 EATTDCKCKGKL-EDTCKKESNCKWENNACKD SSILVTKFAITVVSAAAFVALLF
MITat 1.6 EATPEKCKGDKAK-TCGTQGCWGETCKD SSILVTKFAITVVSAAAFVALLF
ILTat 1.24 DTTEKCKGKYEKDC-KSPDCKWEGGETCKD SSILANKQFALSVSAAAFVALLF
ILTat 1.3 QTDDCKQDKK-DDC-KSPDCKWEGGETCKD SSFILNKQFALSVSAAFAALLF
MVAT 4 DKKEEKCKGKLE-PECTKAECKWGETCKD SSILVDKQFILSMISAAFM...
ANTAT 1.1 ETTPAKTCKKK-DDK-KDGCKWEAETCKD SSILVTKFALSVSAAFAALLF
ANTAT 1.10e EATTDCKDKTK-DEC-KSPNCKWEGGETCKD SSILVTKFALSVSAAFAALLF
ILTat 1.25 EA--KKCSDKKKKEEKC-KSPNCKWDGKECKD SSILANKQFALSVSAAAFVALLF

C-terminal domain type 2

* * * * * * * * * * * * * * * * * * * *

MITat 1.1 LKDVKEKKEKKESETEGKDEK--ANTTGS NSFLIHKA-PLLALFLF
MITat 1.2 KKCNDKKEEAKVADETAKGDGTGNTNTTGS NSFVIKST-PIWIAVLFLF
ILTat 1.22 ECSEMRDGECNAKVAKTAEPSK--NTTGN NSFAIK-TSTLLIALFLF
ILTat 1.21 PPKTGTETATGPOTADPK--ANNTVS DSSLI-KTSPWALFLFLF
ILTat 1.1 EEAUKRVAQFANQETEGKDTGKT--NTTGS NSFVIHKA-PLFLALFLF

C-terminal domain type 3

* * * *** *** *** *** *** *** *** *** ***

MITat 1.5 KENTPGQSAVCGFRIKGDGETDEPDKEKCRN GSFLTSHQFAFSVSAFAFTALLF
ILTat 1.23 EENKGTKTPVCGWGRGREGSDQ-DREMCRN GSFLAKKFKALSVVSAAFTALLF

C-terminal domain type 4

MVAT7 IQNQTBCGKVGGTPTGTAKVCWIRGKCQD SIFLLSKQFALSVSAAFAALLF

Figure 3. The amino acid sequences of the C termini and hydrophobic extensions of VNAs. The gap in the sequence represents the position of the cleavage to form the mature C terminus. The location of the mature C terminus is based on sequence homology to members of the MITAT 1 serotype for which the identity of the mature C terminus has been determined (Holder & Cross 1981). Some gaps (−) have been introduced to increase sequence homology.
terminal domains have a glycosylation site either between the two groups of four cysteine residues (MITat 1-4, MITat 1-6, ILTat 1-24, ILTat 1-2, ANTat 1-1 and ANTat 1-10) or in the region of the boundary between the N and C-terminal domains (MVAT 4 and ILTat 1-25). ANTat 1-1 and ANTat 1-10 have an additional cysteine residue in the region of the domain boundary and MVAT 4 has apparently lost the sixth cysteine residue from the C terminus but has an extra cysteine residue in the domain boundary region; it is not clear from this whether MVAT 4 represents a variant of the type-1 C-terminal domain or a novel structure.

There are five VSGs with a type 2 C-terminal domain. These all have a group of four cysteine residues followed by between 19 and 33 residues before the mature C terminus. The C-terminal residue is serine, except for ILTat 1-22 which has a C-terminal asparagine residue that is also a potential glycosylation site. The hydrophobic extension is 17 residues long and usually starts with asparagine although ILTat 1-21 starts with aspartic acid. All have a glycosylation site four to six residues from the C terminus. The position of cysteine residues relative to the C terminus is more variable than in the type-1 VSGs as is the sequence of the hydrophobic extension. ILTat 1-1 and ILTat 1-22 have an additional cysteine residue towards the region of the boundary between the N and C-terminal domains.

The type-3 C-terminal domain (MITat 1-5 and ILTat 1-23) has some similarities to type 1 in that the domain contains eight cysteine residues. Also the group of four cysteine residues furthest from the C terminus in types 1 and 3 is similar in terms of spacing between cysteine residues. The difference between types 1 and 3 lies in the four cysteine residues closest to the C terminus. The relative positions of the four cysteine residues nearest the C terminus is quite different and the C-terminal residue in type 3 is asparagine, which is also a potential glycosylation site. The hydrophobic extension is 23 residues long and starts glycine-serine-

The C-terminal domain of MVAT 7 is put in a class of its own, type 4. It contains four cysteine residues as does type 2. However, it differs in the following ways: the cysteine nearest to the C terminus is only three residues distant; the C-terminal residue is aspartic acid and the hydrophobic extension is 23 residues in length. There is a glycosylation site 29 residues from the C terminus in MVAT 7 compared with a site four to six residues from the C terminus found in type 2 VSGs.

(c) Classification of VSGs as a combination of domain types

The combination of N-terminal and C-terminal domain found in each VSG is as follows: MITat 1-1, A2; MITat 1-2, A2; MITat 1-5, A5; MITat 1-6, A1; ILTat 1-21, B2; ILTat 1-22, A2; ILTat 1-23, B3; ILTat 1-24, A1 and ILTat 1-25, B1. It is possible to add to this list the domain combinations of previously published VSG cDNAs, MITat 1-4, A1; ILTat 1-1, C2; ILTat 1-3, A1; MVAT4, A1; MVAT7, C4; ANTat 1-1, A1 and ANTat 1-10e, A1. It is immediately apparent that a range of combinations are found. This range is probably only limited by the amount of data available.

4. Discussion

(a) Domain combinations and structure in VSGs

Comparison of known VSG sequences, the new sequences presented here along with previously published sequences, reveals that, in both the N-terminal and C-terminal domain, the VSG domains can be classified based on the pattern of conserved cysteine residues (Fig. 2). The disulfide bonds formed by the cysteine residues in the N-terminal domain have been determined chemically for MITat 1-4 (Allen et al., 1982) and crystallographically for MITat 1-2 (Freymann et al., 1990) and are conserved. The strict geometry constraints necessary to form a disulfide bond imply that conserved disulfide linkages indicate a conserved secondary and tertiary structure (Richardson & Richardson, 1980). The disulfide bond between Cys15 and Cys145 in MITat 1-2 covalently bridges a surface loop to one of the core helices of the N-terminal domain structure. The second disulfide bond lies a strand passing over the top of the domain to a buried part of the chain (Fig. 5 in Freymann et al., 1990). These features can be expected to be present in the other VSG variable domains in which these cysteine residues are conserved, and has been confirmed in a second class-A N-terminal domain (Blum, 1990; M. Blum et al., unpublished results). The other sequence classes have a larger number of conserved cysteine residues. It is not yet possible to rule out that the type-B and type-C N-terminal domain cysteine residues are a superset of the class-A cysteine residues. This is implied by the strict conservation, across all N-terminal classes, of the cysteine residue near position 15, but multiple sequence alignment (Subbiah & Harrison, 1989) using both class A and B sequences (M. Blum, unpublished results) has not established the relationship of the B domains to the class A structure, nor identified the corresponding second cysteine residue of the expected disulfide bond in the class B sequences.

Is there any structural limitation on an individual N-terminal domain limiting the type of C-terminal domain found in the VSG? Could a genetic recombination occur so that the N-terminal domain found in MITat 1-1 (A2) formed a functional VSG with an A1 or A3 domain combination? Genetic recombination is known to occur around the region of the gene encoding the domain boundary, one example occurred between two antigenically distinct members of the same VSG gene family (ANTat 1-1 and ANTat 1-10) with the same domain types (Pays et al., 1988). A second example is that the C-terminal 80 residues of ILTat 1-1 and T. equiperdum VSG 78 (Roth et al., 1986) are of
<table>
<thead>
<tr>
<th>C-terminal domain type 1</th>
<th>α</th>
<th>β</th>
<th>γ</th>
</tr>
</thead>
<tbody>
<tr>
<td>MITat 1-4</td>
<td>ACK 19 RCE 5 ACP 14 RCK</td>
<td>24</td>
<td>NCK 5 RCT 3 GCRN 3 TCKD</td>
</tr>
<tr>
<td>MITat 1-5</td>
<td>ACK 15 RCE 5 GCRN 3 GCKV 14 RCK</td>
<td>27</td>
<td>NCK 5 RCT 3 GCRN 3 TCKD</td>
</tr>
<tr>
<td>ILTat 1-23</td>
<td>ACK 5 RCE 5 GCRN 3 GCKV 14 RCK</td>
<td>27</td>
<td>NCK 5 RCT 3 GCRN 3 TCKD</td>
</tr>
</tbody>
</table>

Figure 4. The relative positions of cysteine residues and conserved flanking residues in the C-terminal domain of VSGs. The standard 1-letter amino acid code is used, the numbers represent the size of the gap in numbers of amino acids not shown. The symbols above the residues indicate (*): cysteine residues; (§): conserved residue in > 60% of examples; (#): conserved aromatic amino acid.

common origin (54/80 identities with gaps introduced), at 80 residues from the C terminus there is an abrupt end to the homology so it is probable that the two VSGs have different N-terminal domains. However, the distribution of glycosylation sites suggests some limitations in that the type-A N-terminal domains have a potential N-glycosylation site in all examples of VSGs of domain combinations A2 (n=2) and A3 (n=1) but only one (ANTat 1-1) from seven examples of A1 VSGs has such a site (Fig. 2). This suggests a difference between the N-terminal domains capable of forming a functional VSG in combination with a type 1, as opposed to types 2 and 3, C-terminal domain.

There are more obvious similarities between the C-terminal domain types. The disulfide linkages are known for the C-terminal domain of one type-1 VSG, MITat 1-4 (Fig. 2(d)). The eight cysteine residues can be divided into two groups of four, each of which shows the same disulfide bond formation with respect to position of the constituent cysteine residues in the primary structure. It has been suggested on the grounds of protease resistance that each of these groups of four cysteine residues represents a tightly folded structural domain (Allen & Gurnett, 1983). Since each of the C-terminal domain types is based on either four or eight conserved cysteine residues and each group of four probably represents a small domain, VSGs can be visualised in T. brucei as a variable N-terminal domain with one (types 2 and 4) or two (types 1 and 3) smaller and relatively conserved domains at the C terminus.

It has been suggested that the type-2 C-terminal domains, which contain four cysteine residues, arose from the type 1 through a deletion event (Cross, 1984). By considering the spacing between the cysteine residues, and the identity of the residues flanking them it is possible to suggest that the groups of four cysteine residues are divided into three patterns, α, β, and γ (Fig. 3). The different C-terminal domain types are derived from unequal crossing overs at the DNA level generating combinations of these groups, type 1, αβ; type 3, αγ; type 2, α except ILTat 1-22, β; type 4, γ. The exception is MVAT 4 which nominally has a type 1 C-terminal domain; however, the first set of four cysteine residues does not match any of the α, β or γ groupings.

It can also be seen in Figure 4 that the identity of the amino acid, two residues to the C-terminal side of the third cysteine in each group of four, is one of the aromatic amino acids (tryptophan, tyrosine or phenylalanine) in 22/34 of the examples shown. It may be significant that the two VSGs that form exceptions to this rule (ILTat 1-22 and ILTat 1-1) are the only two VSGs with a type 2 C-terminal domain with a fifth cysteine residue. The significance of this extra cysteine is clear, in a test of seven VSGs (ILTat 1-1, 1-2, 1-22, 1-23, 1-24, 1-25) it was found that only ILTat 1-22 occurred as a disulfide linked dimer in solution (Miller, 1980). Disulfide linked VSG dimers have also been reported to occur in T. evansi (Cross, 1977). A prediction from this result is that VSG ILTat 1-1 may occur as a disulfide linked dimer in solution. The same func-
tion cannot be attributed to the odd cysteine residues in N-terminal domains as MITat 1-4 does not occur as a disulphide linked dimer (Allen & Gurnett, 1983).

(b) Domain combinations present in each repertoire

There are five characterized VSGs from the MITAT 1 repertoire; all have a type-A N-terminal domain, whereas in the seven characterized members of the ILTAT 1 repertoire, N-terminal domain types A, B and C are found. It could just be chance that all the MITAT 1 VSGs have the same N-terminal domain type; however, there are two other possibilities. Either the MITAT 1 repertoire could be genetically more restricted, or the method of generating the cloned trypanosome populations for each repertoire may have had some effect. The MITAT 1 repertoire was generated by cloning trypanosomes in mice after isolation from a rabbit at different time points after infection, whereas ILTAT 1-21 to ILTAT 1-25 were generated using mice only.

(c) Biological context of VSG domain combinations

The best characterized isolate of T. brucei in terms of sequence information about VSGs is the ILTAT 1 repertoire. Seven of nine VSGs described are sequenced fully and there are sufficient sequence data for the other two to designate their domain combinations with reasonable certainty. ILTAT 1-2, A1 (Rice-Ficht et al., 1981); ILTAT 1-4, A2 (Rice-Ficht et al., 1981; Donelson et al., 1982). The domain combinations can be superimposed on the relationship between the cloned trypanosome stocks from the ILTAT 1 repertoire (Miller & Turner, 1981; Fig. 5). It can be seen in all cases except one that on switching of VSGs, the new VSG has a different domain combination to the old VSG. At first glance, this provides an attractive model for antigenic variation as it would allow the trypanosome a mechanism to distinguish "old" and "new" VSGs. However, this switching from one domain combination to another is not too surprising as among the nine ILTAT 1 VSGs six different domain combinations are found. Furthermore, there are two clear examples where no such switch occurs, ILTAT 1-2 to ILTAT 1-3 in these experiments and ANTAT 1-1 to ANTAT 1-10 (Pays et al., 1983). The above changes in VSG expression were analysed by cloning individual trypanosomes from the population before and after switching. An approach that samples more switching events is needed. Such data are available from more general experiments in which approximately ten trypanosomes from a cloned population were used to infect mice and variants arising in the second peak of parasitaemia were recognized using antisera raised against known variant types (Miller & Turner, 1981). In this way, the percentage of the population corresponding to different domain combinations can be estimated to some extent. The most informative experiments are perhaps those in which trypanosomes expressing ILTAT 1-22 switched to a range of VSGs, but the predominant switch was to ILTAT 1-4. Both ILTAT 1-22 and ILTAT 1-4 are A 2 VSGs, again confounding the hypothesis.

Do the observations above indicate that the domain combination of VSGs has nothing to do with the process of antigenic variation but is merely a result of structural constraints needed to produce a functional VSG? One observation that indicates that there may be more than structural considerations involved is in the identity of VSGs on a dual expressor (that is, a trypanosome expressing more than one VSG at a time). Baltz et al. (1986) described a dual expressing population which arose in culture. However, the two T. congolense VSGs involved are of unknown domain combinations. Some earlier experiments with T. b. rhodesiense (Miller, 1980) involved growing a cloned population expressing ILTAT 1-13 in culture; the trypanosomes were grown for 28 days, then the culture was neutralized with anti-ILTAT 1-13 serum and used to infect a mouse. After four days the VSGs being expressed were characterized by immunofluorescence. All the trypanosomes appeared to be expressing both ILTAT 1-14 and ILTAT 1-22 in roughly equivalent amounts. This population was used to serially infect a second mouse; after a further two days all the trypanosomes were still positive with both antisera but the reaction with anti-ILTAT 1-14 was much stronger than the reaction with anti-ILTAT 1-22. Both ILTAT 1-14 and ILTAT
1-22 are A 2 VSGs, the only two A 2 VSGs from the nine characterized VSGs of the ILTAR 1 serodeme, so the likelihood that the dual expressor has two VSGs of the same domain combination by chance is small. These results suggest that a period in culture occasionally leads to expression of two VSGs by one trypanosome and that if the two VSGs are of the same domain combination then the dual expression is stable for a few days when transferred to an animal. Trypanosomes expressing two VSGs of different domain combinations are either unstable in culture or in the animal, and rapidly switch to a trypanosome expressing one or the other but not both. Under the normal in vivo conditions for antigenic variation, a rapid switch is desirable. Therefore variants in which a switch in domain type has occurred would be at an advantage over those in which the domain combination remained the same.

We thank Linda Allen and Paul Vorhees for help in preparing bloodstream-form trypanosomes. This work was funded by Merek Sharp & Dohme and the Wellcome Trust. The oligonucleotide synthesis facility in the Biochemistry Department in Cambridge is supported by the Wellcome Trust.

References


Edited by J. Kara