A structural motif in the variant surface glycoproteins of *Trypanosoma brucei*

Michael L. Blum*†, James A. Down*†, Anne M. Gurnett‡, Mark Carrington§, Mervyn J. Turner‡ & Don C. Wiley*‖

‖ Howard Hughes Medical Institute, and * Department of Biochemistry and Molecular Biology, Harvard University, 7 Divinity Avenue, Cambridge, Massachusetts 02138, USA
‡ Merck Sharpe & Dohme, Research Laboratories, Box 2000, Rahway, New Jersey 07065, USA
§ Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QW, UK

The variable domain of the trypanosome variant surface glycoprotein (VSG) ILTat 1.24 has been shown by X-ray crystallography to resemble closely the structures of VSG MITat 1.2, despite their low sequence similarity. Specific structural features of these VSGs, including substitution of carbohydrate for an α-helix, can be found in other VSG sequences. Thus antigenic variation in trypanosomes is accomplished by sequence variation, not gross structural alteration; the extensive sequence differences among VSGs may be required for another reason, such as the avoidance of recognition by helper T cells. Additionally, VSG sequences are found to define families, within a VSG superfamily, which have evolved in the trypanosome genome.

THE African trypanosome is a parasitic protozoan that causes sleeping sickness in humans and nagana in cattle, diseases that are characterized by cyclical waves of fever. The fever cycles correlate with spikes in the population of trypanosomes in the blood[1]. The trypanosome uses a large, serially expressed repertoire of antigenically distinct surface proteins to evade the immune response of its mammalian host[2,3]. These antigens, the variant surface glycoproteins, form a coat on the surface of the parasite[4,5] and elicit the production of VSG-specific host antibodies[5]. Some of the trypanosome population survives the host's immune response by switching to expression of one or more antigenically distinct VSGs, allowing a relapse in parasitaemia. The trypanosome genome is estimated to contain over 1,000 different VSG genes sharing little sequence homology[6]. No biochemical activity has been ascribed to these molecules, leaving the ability to fold and form a coat as the only known common requirement for their sequences.

Most VSGs consist of two domains[7]. The N-terminal 'variable' domains share low sequence homology (13–30% identity)[8,9], contain the antigenic epitopes accessible on the surface of the parasite[10], and have recently been grouped by cysteine patterns into three classes A, B and C[8]. The C-terminal, 'conserved' domains are grouped by sequence homology into three classes, I, II and III (refs 8, 11), which appear to pair randomly with the three N-terminal classes in known VSGs. The three-dimensional structures of two class A, N-terminal domains with low sequence similarity have been determined by X-ray crystallography, namely MITat 1.2 (refs 12, 13), and ILTat 1.24 reported here[14] (for VSG nomenclature, see legend to Fig. 2).

Similarities in the two three-dimensional structures of VSGs and weak amino-acid homologies, identified only by multiple sequence alignments with other class A N-terminal sequences, are here correlated to define a folding motif common to most VSGs. This analysis allows specific predictions of intra- and interchain disulphide bonds and missing secondary structure elements in other VSG sequences. It argues that trypanosomes accomplish antigenic variation by sequence variation and not by major structural alterations. We also observe that carbohydrate can substitute for deleted polypeptide in the MITat 1.2 VSG tertiary structure. This type of substitution can be inferred in another VSG sequence, suggesting that it represents a novel structural role for oligosaccharide. The structural similarities of VSGs raise questions about the function of what appears to be a great excess of sequence diversity relative to that needed to escape antisera. They also suggest that the VSG variable domain classes (A, B, C) represent a protein superfamily developing within one genome as the result of a diversity generator, presumably evolved for antigenic variation.

Structure determination

The structure of the N-terminal domain of the ILTat 1.24 VSG was determined by X-ray crystallography. A single isomorphous heavy-atom derivative (HgI4) and its anomalous scattering provided phases to 4.2 Å resolution, which were extended to 3.8 Å resolution by solvent flattening and iterative, real-space, non-crystallographic symmetry averaging about a molecular twofold axis[15]. Fifty per cent of the atomic model was readily constructed, including correction of a sequence error (a 3-residue insertion). Four cycles of model building, refinement, phase combination and non-crystallographic phase averaging were required to complete the atomic model, which is currently refined at 2.7 Å resolution to an R factor of 20.3% with good geometry (r.m.s. $\Delta_{bonds} = 0.019$ Å, r.m.s. $\Delta_{angles} = 3.47°$) (see Fig. 1 legend). An indication of the quality of the electron density maps is that three sequence errors were discovered (one insertion and two phase shifts in the nucleic acid sequence) by interpreting the electron density, and were later confirmed by resequencing[9].

Structure of ILTat 1.24

The central elements of the ILTat 1.24 variable domain are two antiparallel α-helices (A and B in Fig. 2a) with a short turn (c) between them. From this scaffold are hung various smaller elements of secondary structure: 7 smaller α-helices (C, D, E_0, E, F, H, S) and a short, three-stranded β-sheet (near 'n' in Fig. 2). These elements are tied together by several loops (e, n, o, p, q, r, s, t, u) and one long meandering strand that covers much of the upper side (h to k) and top (k to m) of the molecule (Fig. 2, left). The N terminus is at the top of the molecule (a) and the C terminus exits at the bottom (v), presumably where it

† Present addresses: Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA (M.L.B.); Beckton Dickinson and Company, Research Center, Box 12016, Research Triangle Park, North Carolina 27709, USA (J.A.D.).

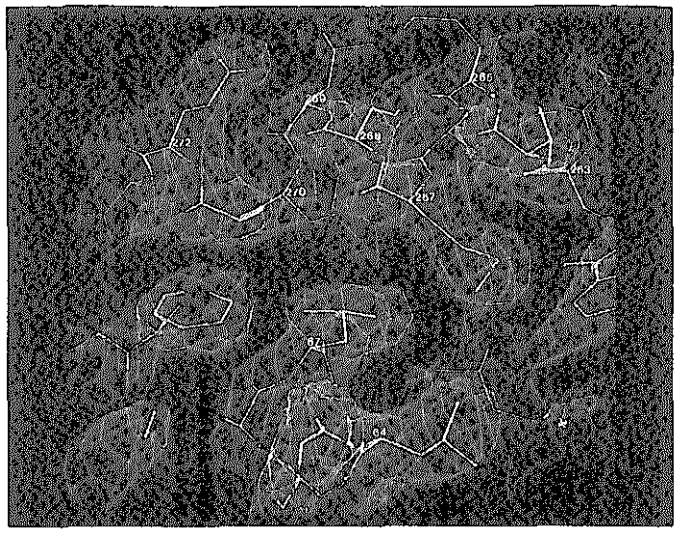

FIG. 1 Electron density map of ILTat 1.24 and the current atomic model. Electron density was calculated using current model phases and $2Fo-Fc$ terms between 10 and 2.7 Å. The view is looking up toward the E_o helix and contacts with the B helix. Initial crystallographic phase determination and identification of the non-crystallographic symmetry (NCS) axis were carried out at 6.0 Å (ref. 47). Phases to 4.2 Å were calculated using single isomorphous replacement differences and anomolous scattering from a mercury (K_2HgI_4) derivative and improved by iterative NCS averaging and solvent flattening[15]. Native data to 2.7 Å were collected from seven crystals yielding 24,682 unique reflections from 107,903 independent measurements with an overall R_{sym} of 0.102. Lack of crystals prevented collection of useful high-resolution derivative data. Phases were extended to 3.8 Å by continued iteration of NCS real space averaging and phase combination. The low symmetry (twofold) prevented further extension. A partial atomic model, incorporating ~50% of the expected number of atoms, was built, using FRODO (ref. 48), into the 3.8 Å map. Subsequent refinement using CORELS and X-PLOR, phase combination with the partial model, and continued iterative application of NCS real space symmetry averaging allowed completion of an atomic model at 2.7 Å resolution. The final stages of refinement required relaxation of the NCS constraints owing to a significant bend in the molecular dimer axis. The current model consists of 5,462 atoms, including 90 H_2O molecules with mean $B=27$ Å^2, $\Delta_{bonds}=0.019$ Å and $\Delta_{angles}=3.47°$, and $R=20.3\%$ for 21,676 reflections ($F>2\sigma_F$) between 6.0 and 2.7 Å (details of the method of structure determination will be presented elsewhere). Analysis of solvent accessibility and structural congruence to MITat 1.2 presented in Fig. 4 was on the NCS constrained model[14]. Relaxation of this constraint does not significantly alter the discussion.

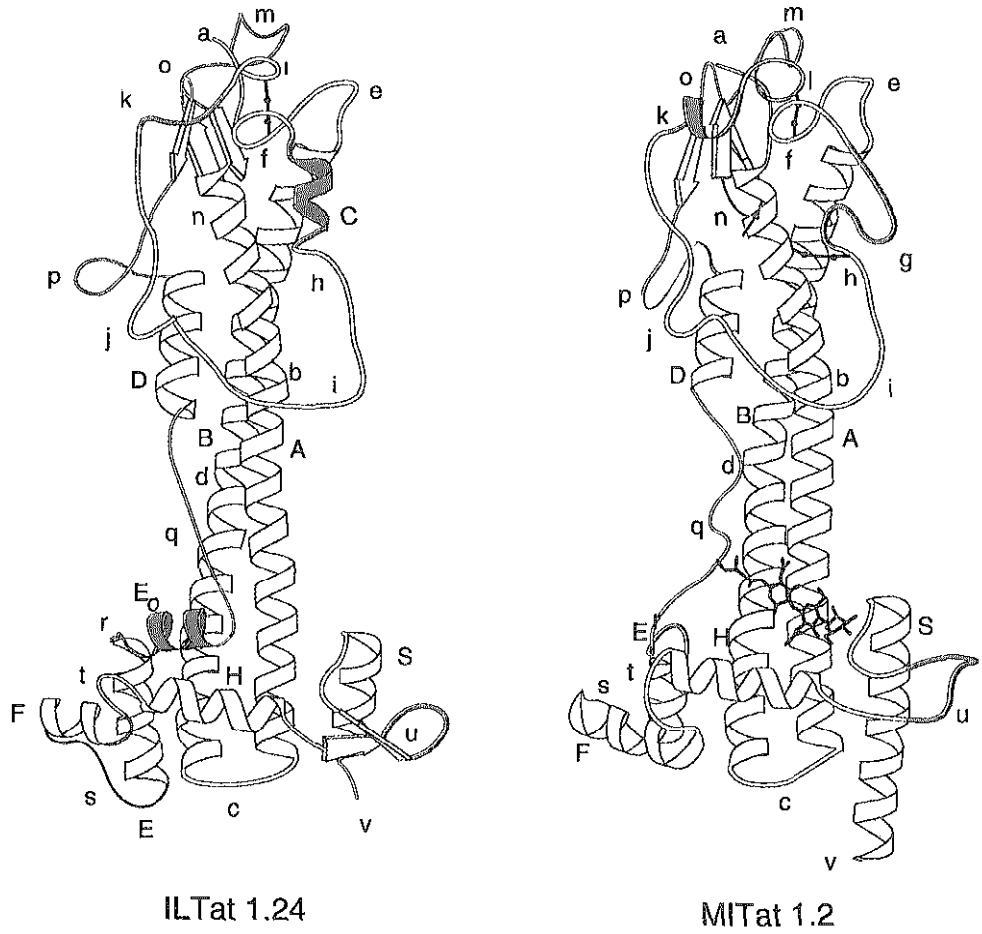

FIG. 2 Ribbon drawings[49,50] of the two known VSG structures MITat 1.2 and ILTat 1.24. Coloured regions highlight the principal structural differences between the two molecules. Helices are identified by upper-case letters, other features are labelled with lower-case letters. a, N terminus; b, kink in A helix; c, turn between A and B helices; d, kink in B helix (MITat 1.2 only); e, loop at top of B helix; f, buried loop with disulphide bond; g, small β-sheet (MITat 1.2 only); h, loop with disulphide bond to A helix; i, large side loop; j, conserved hairpin turn; l, extended strand across top; m, dimer contact loop; n, turn one of β-sheet; o, turn two of β-sheet; p, loop at top of D helix; q, strand connecting D helix to E (MITat 1.2) or E_o (ILTat 1.24) helix; r, E_o/E bend (ILTat 1.24) or CHO attachment site (MITat 1.2); s, E/F loop; t, F/H loop; u, long loop between H and S helices; v, C terminus. MITat 1.2, Molteno Institute *Trypanozoon* antigen type 1.2; ILTat 1.24, International Laboratory for Research on Animal Diseases, Trypanozoon antigen type 1.24.

enters the C-terminal domain, which was removed by an unidentified proteolytic activity before crystallization[7,16]. ILTat 1.24 is a dimer formed by the close association of the long helices of the monomers into a 4-helix bundle, much the same as that previously determined for MITat 1.2 (ref. 13).

Structure shared with MITat 1.2

Overall, the structure of ILTat 1.24 and MITat 1.2 appear similar (Fig. 2). Using an objective measure of the similarity, which considers the spatial arrangement of corresponding residues and their orientation with respect to the preceding and succeeding residues[17], 207 residues (60%) are found to form the same structure. This shared core of structure, coloured in Fig. 3, is comprised of the two long helixes coloured blue, five smaller helices that decorate the central helices coloured red, parts of three β-strands and a buried turn between two strands coloured orange, and five short segments of loop coloured yellow. The 207 residues of common structure share only 16% sequence identity (Fig. 4) and have an r.m.s. deviation in Cα positions of 1.8 Å, similar to that found in other families of protein structure[18]. When aligned by sequence alone, the entire N-terminal domain sequences of MITat 1.2 and ILTat 1.24 show a higher (20%) sequence identity, but this increased similarity is an artefact of sequence alignment unconstrained by knowledge of tertiary structure. The remaining 40% of the structure of the two VSGs is not conserved in detail, but within this portion there are only six segments of significantly different structure (coloured red in Fig. 2). Two α-helices in ILTat 1.24 (C and E_0) are strands in MITat 1.2, one α-helix in MITat 1.2(k) is a strand in ILTat 1.24. Three loops (m, p, u) also have different conformations.

One region of different structure is particularly interesting because it includes the single *N*-linked glycosylation site in the MITat 1.2 variable domain sequence. The first three monosaccharide residues at this site were sufficiently well ordered in the MITat 1.2 electron density map that they could be modelled[14]. The E_0 helix in ILTat 1.24 is found to occupy about the same volume in space as that occupied by carbohydrate in MITat 1.2 (Fig. 2). Amino-acid positions in the conserved core that contact monosaccharide residues in MITat 1.2 contact amino-acid residues of the E_0 helix in ILTat 1.24 (not shown). Thus an oligosaccharide in MITat 1.2 appears to substitute for a helix in ILTat 1.24, each burying residues of the conserved core.

Domain class evident in complete sequence

Although sequence homologies have been noted in the first 30 amino-acid residues of VSGs[19], analysis of complete sequences has generally not identified a basis for a common VSG structure[20–22]. A recent attempt at aligning all of the known complete VSG sequences[22] used pairwise sequence alignments, a method known to be inadequate on sequences of such low homology[23]. In the absence of added structural information, such alignments are no more likely to be correct than many other possibilities[23]. (The alignments in question[22] even failed to match conserved cysteine positions.) The following analysis corrects problems with earlier alignment attempts in light of the realization that differences among all VSGs are not uniformally distributed and that VSG sequences fall into identifiable families. The known three-dimensional structures of two VSGs provide powerful clues for verification of the new alignments and provide a context in which to evaluate the role of observed sequence homologies. Recent identification of variable domain classes based on cysteine positions[8] suggested that sequence analysis might benefit from prior segregation based on this classification. To test this, twenty-three VSG variable domain sequences were aligned, pairwise, using a standard algorithm[24] and a protein similarity scoring matrix[25]. When the sequences are divided into classes (A, B, C), they are found to share significant homology among members of a class, and to show no statistically significant homology between classes. The range of individual scores was low and wide (from -1.6 to 26.7σ, mean of 2.8), but a sequence could be classified by score alone when its mean score was calculated for all sequences of each class (Table 1). This indicates that classes are evident in the entire sequence, not only in the cysteine positions. The sequences, segregated by class and then aligned within each class, further indicate that a majority of secondary and tertiary structure must be conserved within classes.

Structural motif in class A sequences

Segregation of the VSG sequences by class greatly facilitated alignment (Fig. 4) and further suggested that features of the aligned set of sequences might be correlated with structural features found to be conserved between MITat 1.2 and ILTat 1.24. In an alignment of ten class A, VSG variable domain sequences, including MITat 1.2 and ILTat 1.24 (Fig. 4), 81 positions were found to exhibit conservation of the character of the amino-acid residue across eight of the ten sequences (blue

TABLE 1 Mean alignment scores for 21 VSG N-terminal sequences when aligned pairwise with all other VSG N-terminal sequences

Variant	All As	All Bs	All Cs
MVAT 5	5.7	0.8	1.8
AnTat 1.10	5.6	0.9	2.3
ILTat 1.3	7.0	1.0	0.2
MVAT 4	4.4	0.7	1.2
MITat 1.1	8.1	1.0	2.1
MITat 1.4	6.7	0.7	2.6
MITat 1.5	4.1	1.1	0.4
MITat 1.6	7.0	1.1	1.1
ILTat 1.22	1.7	0.0	3.2
ILTat 1.24	7.1	1.0	1.8
MITat 1.2	8.5	1.1	3.2
All class A	6.0	0.9	1.8
ILTat 1.21	0.4	7.1	0.6
ILTat 1.23	1.1	4.4	1.5
ILTat 1.25	1.8	5.3	0.8
WRATat A	1.8	7.3	0.2
WRATat B	0.5	6.5	0.5
YNat 1.1	0.3	4.5	0.4
YNat 1.3	0.0	5.6	−1.1
All class B		5.8	0.4
BoTat 20	2.4	−0.4	10.1
ILTat 1.1	1.8	0.3	11.5
MVAT 7	1.3	1.3	12.4
All class C			11.4

The program ALIGN[24,40] and a similarity scoring matrix[25] were used. The true sequence length of the N-terminal domain was approximated by truncating the sequence one residue short of the first cysteine residue recognized as belonging to the C-terminal homology domain. Scores are expressed as number of standard deviations above the mean score for 50 alignments of randomized sequences. The sequences represent all the published sequences (as of November 1991), excluding those that are very closely related (for example, AnTat 1.1, which is 70% homologous to AnTat 1.10). Each VSG sequence, when aligned pairwise with all other sequences, scores best against other sequences of its class. ILTat 1.22 is an exception, scoring better against the class C sequences than against the other As. The cysteine positions in ILTat 1.22 do not appear to be related to the class C cysteine positions; it is possible that it is either a hybrid or a member of new class. The intraclass scores are significantly greater (by 4 to 12 σ) than random. The interclass scores are only barely (1 to 2σ) significant. The mean score was 3.0 for all 210 possible pairwise alignments. The sequences were classified as A, B or C based on the pattern of cysteines[8]. Class A: MVAT 5 (ref. 41), AnTat 1.10 (ref. 42), ILTat 1.3 (ref. 9), MVAT 4 (ref. 43), MITat 1.1 (ref. 8), MITat 1.4 (ref. 44), MITat 1.5, MITat 1.6, ILTat 1.22, ILTat 1.24, and MITat 1.2 (ref. 8); class B: ILTat 1.21, ILTat 1.23 and ILTat 1.25 (ref. 8), WRATat A, and WRATat B (ref. 41), YNat 1.1 and YNat 1.3 (ref. 21); class C: BoTat 20 (ref. 45), ILTat 1.1 (ref. 46), MVAT 7 (ref. 43).

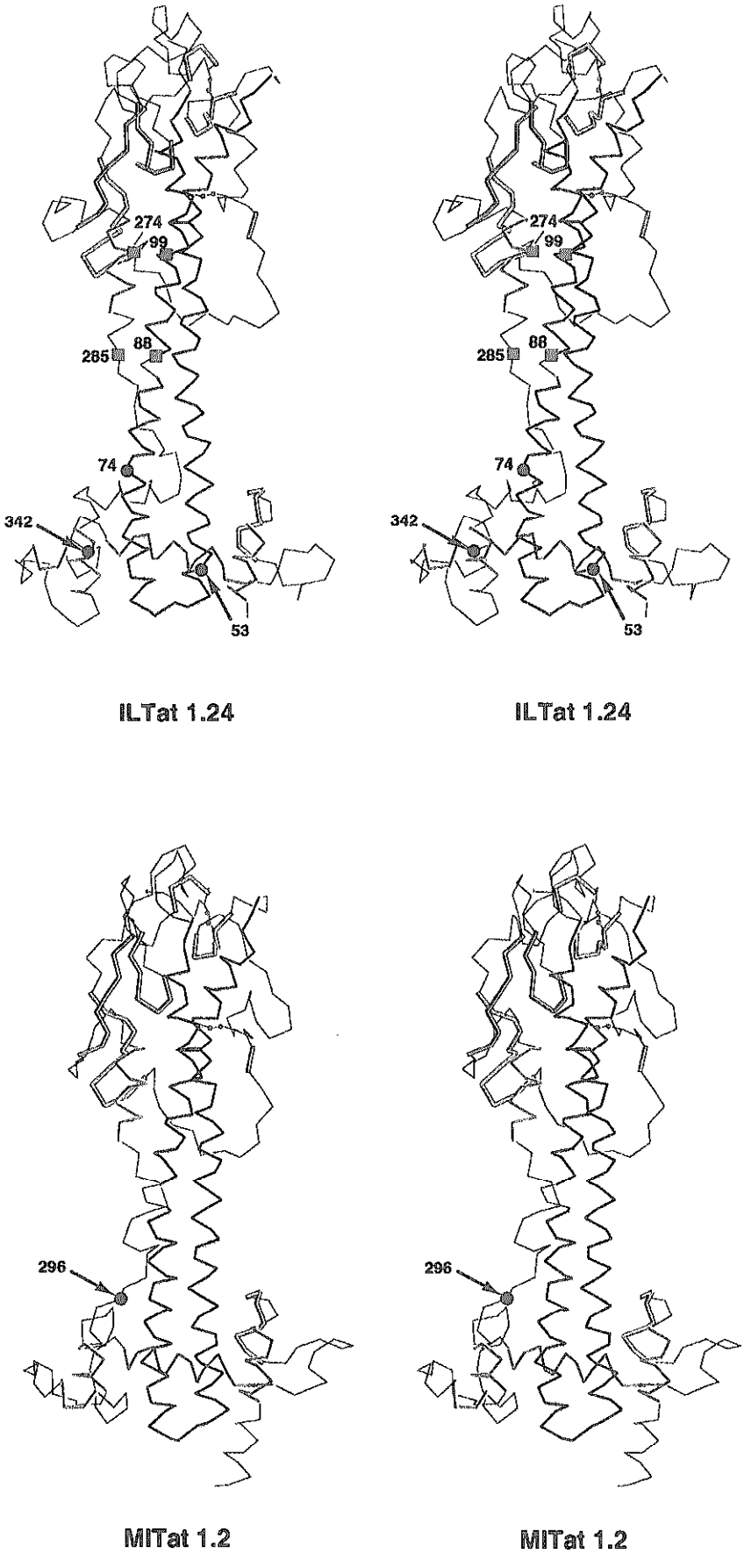

FIG. 3 Stereo drawings of trace through Cα positions of the two known VSG structures. The conserved core structure (see text) is coloured to correspond to sequence alignment in Fig. 4. Dark blue, central helical core; yellow, various conserved loops; orange, β-sheet region; pink, smaller conserved helices. Top, ILTat 1.24. Black dots show where glycosylation sites in MITat 1.5 map to the ILTat 1.24 structure based on sequence alignment (Fig. 4) These sites lie such that carbohydrate at 53 and 345 and at the dimer-related 74 would be positioned where they might substitute for parts of the H or S helix (see text). Bottom, MITat 1.2. The observed glycosylation site at 296 in MITat 1.2 is marked. Note that sequence numbers refer to the consensus sequence of Fig. 4 and not to the individual molecular sequences. Red squares indicate the positions of cysteines in the sequences of MVAT 4 (99 and 274) and MITat 1.6 (88 and 285) based on the sequence alignment. Each of these pairs is predicted, by their proximity in the folded structure, to form a disulphide bond.

columns in Fig. 4). Three-quarters of these positions are buried (inaccessible to solvent) in both of the VSG structures (unpublished data), which suggests that their character has been conserved to maintain intramolecular contacts that stabilize the VSG fold. Most significantly, the positions of the eighty-one conserved amino acids cluster in regions (black blocks, Fig. 4) where MITat 1.2 and ILTat 1.24 share common structure (coloured rows in Fig. 4) and display patterns indicative of the observed secondary and tertiary structure.

Specific features of the shared core structure are evident in the pattern of conservation in the aligned, class A sequences. Quasi-periodic repeats[26] suggest the presence of the A, B and D helices (dark blue rows in Fig. 4) and conserved residue character is present in all the other helices (violet blocks in Fig. 4), except F, which has only 1.5 α-helical turns conserved between ILTat 1.24 and MITat 1.2. Conserved residue character is also present in the three-stranded β-sheet (orange blocks, Fig. 4), and four of the five conserved loop segments (yellow blocks, Fig. 4). Regions of conserved glycine or proline (green columns, Fig. 4) define several turns; appearing between the two long

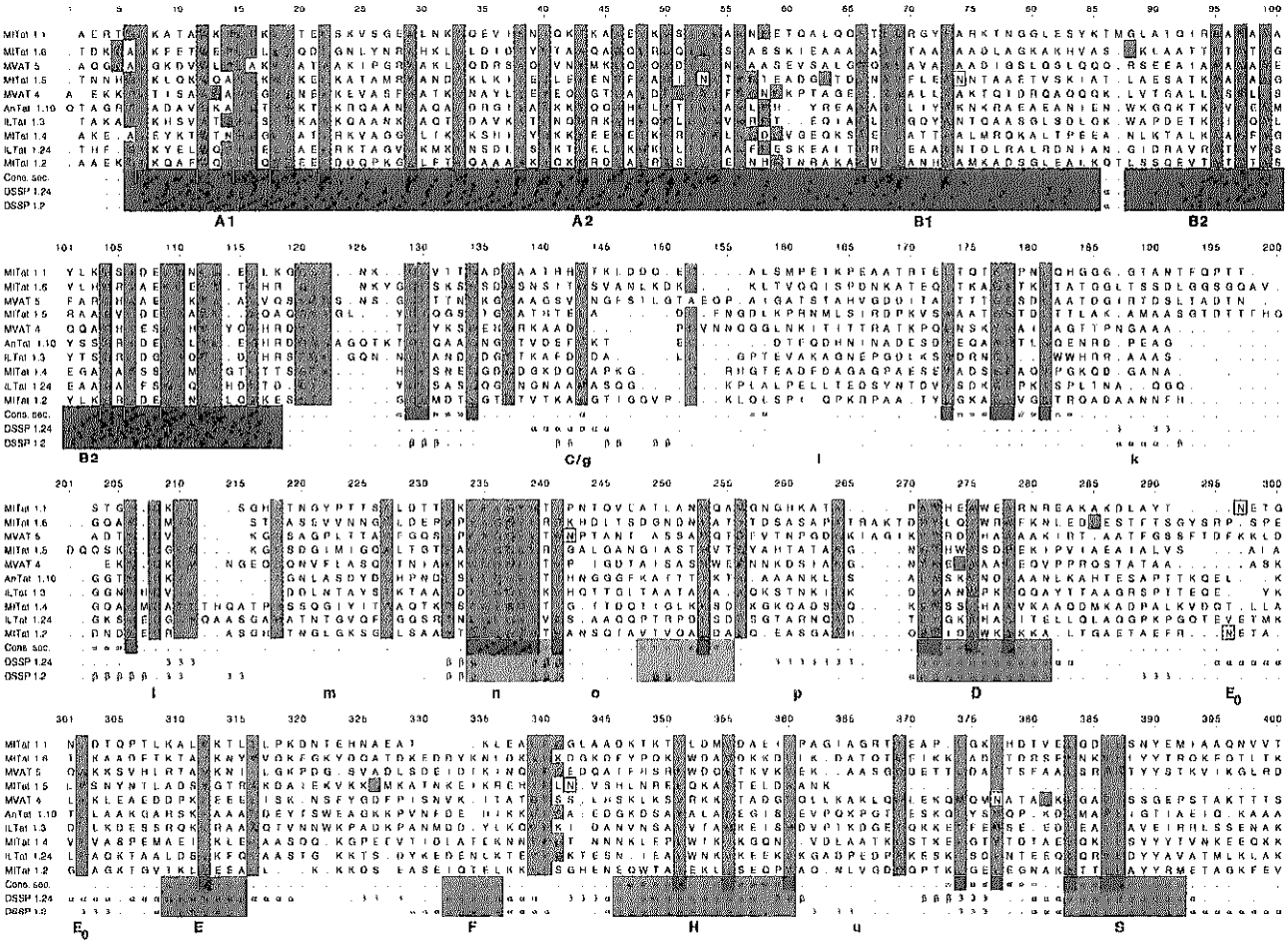

FIG. 4 VSG sequences alignments. The aligned sequences of 10 class A VSG variable domains showing the relationship of conserved structural features of ILTat 1.24 and MITat 1.2 to identifiable homologies in the aligned sequence set. Alignment was accomplished in two steps. First, all 10 of the sequences were aligned with an *n*-dimensional alignment computer program (TULLA[51]) using the protein similarity scoring matrix[26]. Then, the sequences were adjusted manually (as recommended by S. Subbiah, personal communication) using a sequence editor (LINEUP[52]) in order to align recognizable features, such as the region between 232 and 241, that the program did not successfully align across all sequences. The failure of the program to correctly align these features was largely due to large deletions in nearby regions in one or more sequences. Extensive manual adjustment was also necessary to reconcile the alignment with the known three-dimensional structural alignment between ILTat 1.24 and MITat 1.2. The alignment of these two sequences, presented here, is based on the known three-dimensional structures. Adjustment of the alignment using knowledge of the three-dimensional structures[23] was crucial to recognizing structural features which are indeed present in the properly aligned sequences, but are hidden by the low sequence homology. The use of structural information identified the conserved β-sheet region (232–241), and the hydrophobic conservation in helix H, that have been missed in earlier alignment attempts[22]. Three lines below the sequences indicate structural properties: Cons. Sec., an 'equals' sign is shown at each position where ILTat 1.24 and MITat 1.2 have similar secondary structure as determined by OVRLAP[17]. DSSP 1.24 and DSSP 1.2, secondary structure designation as determined by the program DSSP[53] for ILTat 1.24 and MITat 1.2 respectively where α indicates α-helix, β indicates β-sheet and 3 indicates 3/10-helix. Horizontal coloured blocks identify regions of conserved structure between ILTat 1.24 and MITat 1.2 and correspond to the colour coding of Fig. 2. Sequence alignment is highlighted as follows: light blue, homologous residue in 8 of the 10 sequences; red, cysteines; green, glycine or proline in positions defining conserved turns; boxed residues are potential *N*-linked glycosylation sites. Black rectangles indicate positions where class A sequence homology falls in a region of observed structure conservation. Bold faced lettering below the alignment indicates structural features as labeled in Fig. 2. Homology is defined liberally as hydrophobic, Y=W=F=M=L=I=V=A=G; polar, N=Q=T=S=G; positive charged, K=R; and negative charged, E=D. The figure is truncated at position 400 (out of 425) due to space considerations. An alignment of class B sequences, where no X-ray structure is yet known, is available from the authors. It suggests that class B VSGs contain two disulphides homologous to those conserved in class A sequences. The predicted disulphides are, for example, ILTat 1.25 cysteine 177 to 240 and 16 to 198 or 200. Because of the low sequence homology, the alignments of any one sequence in both class A and more so in class B lists may still contain errors in the details, but such errors are not expected to change the overall picture that emerges from them collectively.

helices (57–59), at the end of helix B (120), at a reverse turn in segment j (177), at the turn between strand 1 and 2 of the three-stranded β-sheet (237), and between helix F and H (341). A conserved glycine at position 104 (Fig. 4) defines a position in the VSG dimer interface where two B2 helices are able to make their closest approach[13].

The four cysteine residues present in both MITat 1.2 and ILTat 1.24 (positions 16 and 152, and positions 129 and 209 in Fig. 4). are the defining characteristic of class A sequences[8], and form two disulphide bonds which are known to be conserved in the two structures (Figs 2 and 3) and in MITat 1.4 (ref. 27). The MVAT 5 sequence lacks both members of the 16,152 disulphide pair (Fig. 4), but fits the remainder of the conserved core pattern (blue columns in Fig. 4), arguing that the disulphide bonds are not absolutely required to maintain the folding motif. The bonding of additional cysteines present in some of the class A sequences can be inferred when the sequences are folded into the class A motif, and provide strong evidence that the alignment is valid. Both MVAT 4 and MITat 1.6 have pairs of cysteines (99, 274 and 88, 285 respectively) which are predicted by the alignment to occur in ideal juxtaposition to form disulphide bonds (Fig. 3). Antat 1.1 and 1.10 both have a single, odd cysteine that maps to the end of the S helix (position 410; not shown in Fig. 4) where a disulphide could form between monomers in the dimer and both molecules have been observed to occur as disulphide-linked dimers[28]. MITat 1.4 also has an odd cysteine but at 275, a location mapping to helix D, that could not possibly form a symmetric, inter-monomeric disulphide bond. That cysteine has been observed to be a free sulphydryl[27].

Carbohydrate can be inferred to substitute for a region of polypeptide in the MITat 1.5 class A variable domain. Six of the seven potential *N*-linked oligosaccharide attachment sites in the sequences in Fig. 4 appear from the alignment with the known structures to be on the surface of the VSG. MITat 1.5 contains three glycosylation sites (53, 74, and 342) but one site (53) would be buried by contacts with the H helix (Fig. 3, top). In the aligned sequences (Fig. 4), the H and S helices are characterized by periodic homologous residues (blue columns) coinciding with buried positions in both MITat 1.2 and ILTat 1.24, a characteristic of amphipathic helices with a hydrophobic face to the interior. The A class VSGs all share these features, except MITat 1.5, which has a shorter sequence apparently lacking all, or most, of the H and S α-helices. Residues 52 and 69 on the A and B α-helices are conserved hydrophobic positions buried by the H and S α-helices. The cluster of three oligosaccharide sites in MITat 1.5 (Fig. 3, top, black dots) facing toward the volume occupied by the H and S helices in the VSG dimer suggests that carbohydrate may bury these conserved core residues in the MITat 1.5 structure, effectively substituting for polypeptide just as carbohydrate in MITat 1.2 appears to replace the E_0 α-helix in ILTat 1.24 (Fig. 2).

Other VSG classes

The patterns of conservation of cysteines and of sequence homology, particularly of hydrophobic residues, in other VSG sequence classes can be used to extend structural arguments to those classes. The variable domain class B currently has a sufficient number of known sequences that patterns of conserved homology can be identified. All known VSG sequences in class A, B and C (except MVAT 5) have a cysteine near position 13–16. This in itself makes some conservation of tertiary structure likely but no evidence is yet available to indicate the oxidation state or possible disulphide partner of this cysteine in any but the class A proteins. Comparison of a set of seven class B sequences (data not shown, but available from authors) reveals that these sequences have a pattern of conserved residue type in the region surrounding the first cysteine that is similar to the pattern observed in class A (Fig. 4). The class B sequences also have in common with class A sequences, two long regions of reduced Pro and Gly content, with hydrophobic heptad repeats, separated by a four-position region of high Pro and Gly content. This indicates that the class B structures are based on two long antiparallel α-helices connected by a turn as observed in the class As, although the length of the class B α-helices and the exact form of the turn may vary. The class B sequence alignment also shows several highly conserved hydrophobic heptads near the C-terminal end of the variable domain (not shown), indicating 3 or 4 medium length helices as in the class A structure, and a few conserved glycine positions indicating conserved turns. As yet, we have not succeeded in matching these features to specific features in the class A structures, although the alignment does suggest a probable disulphide bonding pattern resembling the class A pattern (Fig. 4 legend). There are not yet enough sequences available to analyse the class C VSGs.

Discussion

Our structural studies of ILTat 1.24 and MITat 1.2 VSGs demonstrate that the African trypanosomal antigens accomplish antigenic variation through variation of sequence and limited conformational modification and not by gross alteration of structure. The two structures are essentially superimposable over 60% of their residues, although their amino-acid sequences share only 16% identities in those regions. Using an *n*-dimensional alignment algorithm supplemented with manual adjustments and knowledge of the structural overlap of ILTat 1.24 and MITat 1.2, ten class-A VSG sequences were aligned, making it possible to identify a set of positions where amino-acid character (but not identity) is conserved. Most of these positions are buried in the known structures and therefore contribute to the stability of the VSG fold. The set of these conserved positions overlaps almost all the shared features of secondary and tertiary structure in the two known structures. This argues that all of the class A sequences have elements of this common three-dimensional structure, presumably with some differences like those seen between MITat 1.2 and ILTat 1.24 (Fig. 2). From the sequence alignment, specific details of the structure of unknown VSGs have been inferred, for example the positions of intra- and interchain disulphide bonds, the location of oligosaccharide sites and, in some cases, the absence of elements of structure. A novel role for carbohydrate may have been discovered: an oligosaccharide is observed in MITat 1.2 to substitute for the deletion of an α-helix in ILTat 1.24 by burying the same core residues that the polypeptide had buried; a similar substitution has been inferred from the sequence of MITat 1.5, in which a cluster of three oligosaccharides appear to be positioned to replace part of a surface structural element consisting of two helices and a loop between them that seems to be deleted in that sequence.

The diversity present in VSG sequences appears to be much greater than is required to alter antibody-binding sites on VSGs in order to escape neutralization. In the case of viruses like influenza and rhinovirus, antigenic variation sufficient to escape neutralization from antibodies raised in previous infections requires only about a dozen point mutations on the surface of exposed proteins[29,30]. Point mutations in VSGs also alter antibody binding to exposed epitopes[31]. Yet essentially the entire VSG sequence, comprising most of the surface and most of the interior of the N-terminal domain of the protein is changed between trypanosome antigenic variants. Why does the trypanosome VSG exhibits such extreme sequence variation? Two answers seem plausible. First, the trypanosomes' diversity generator[32–34] may simply be so powerful that it has generated more diversity than necessary to escape neutralization. Excess sequence diversity, beyond that needed for antigenic variation, may simply be accommodated to the limit that will still fold into a VSG structure and form a cellular coat. A second possibility is that the trypanosome requires the VSG sequence diversity to escape recognition both by antisera and by helper T cells.

Helper T cells recognize short peptides processed from antigens and presented on host major histocompatibility complex (MHC) molecules. To escape helper T-cell recognition might require variation throughout the VSG antigen's sequence in order to alter all potential peptides. Effective T-cell help during a viral infection seems to be short-lived, peaking in mice within 3 to 9 days of initial infection[35]. Re-infection within 4 or 8 days with a serologically distinct viral strain, which nevertheless shares helper epitopes, is rapidly neutralized, but re-infection after ≥15 days is not. Reinfection by new virus strains of previously infected individuals often occurs only after weeks, months or years, so that antigenically distinct viral strains would need only to be serologically distinct and not variable at the level of T-helper epitopes (HIV-1 infection may be an exception[36]). The timing of a secondary trypanosome infection by an antigenic variant in trypanosomiasis, however, occurs within several days of the first infection[1]. Trypanosomes may require variation in the VSG helper epitopes, achieved by changes in residues over the entire VSG sequence in order to avoid the T-cell help available within the effective lifetime of 4–8 days. T-cell help plays a part in increased antibody response to VSGs[37], but the nature and importance of the helper epitopes during successive waves of parasitaemia has not been established.

Structural arguments suggests that the three classes of N-terminal VSG sequences, A, B and C, comprising the thousand VSG genes in a trypanosome, may represent a developing protein superfamily. Only one N-terminal proximal cysteine is common to all sequences, but where a number of sequences can be aligned (class B), periodic conservation of mostly hydrophobic residues divided by a glycine-rich turn region argues for the presence of the two long, antiparallel α-helices (A and B) found at the core of the class A structures. Determination of the disulphide bond pattern or the three-dimensional structure of class B and C VSGs may be required to further define the relationship (a prediction is given in Fig. 4 legend). But if, as seems likely, all VSGs have recognizably similar structures, the disparate VSG classes, like the antibodies, CD4, CD8, MHC antigens, T-cell receptors and ICAMs of the immunoglobulin superfamily[38], may represent a protein superfamily whose membership is more difficult to diagnose from sequence alone. Evolution of a protein superfamily evidently requires the duplication of genes and subsequent (segregated) divergence, but in trypanosomes even the duplicated copies of genes appear to be linked through gene conversion events[39]. Class A, B and C genes may be able to mix together in diversity-generating recombinations (or gene conversions), but A, B and C classes have sufficiently diverged that proteins resulting from mixing classes would either not fold stably or not form VSG surface coats. Alternatively, the groups of A, B and C genes may have become segregated by some genetic mechanism so that the diversity-generating mechanism does not mix them. In either case, if sequenced, the thousand VSG genes may offer the opportunity in one genome for analysing the evolution of superfamily-level diversity in a protein fold.

Whatever the cause, the extreme sequence variation of VSGs observable within the trypanosome genome may provide important clues about what sequence constraints are imposed by a particular protein fold. The ability to align the VSGs should greatly facilitate this study of VSG structure. □

Received 9 March; accepted 16 March 1993.

1. Ross, R., & Thomson, D. *Proc. R. Soc. Lond.* **B82,** 411-415 (1910).
2. Donelson, J. E. *Contrib. Microbiol. Immun.* **8,** 138-175 (1987).
3. Cross, G. A. M. *A. Rev. Imnun.* **8,** 83-110 (1990).
4. Vickerman, K. *J. Cell Sci.* **5,** 163-193 (1969).
5. Cross, G. A. M. *Parasitology* **71,** 393-417 (1975).
6. Van der Ploeg, L. H. T. *et al. Nucleic Acids Res.* **10,** 5905-5923 (1982).
7. Johnson, J. G. & Cross, G. A. M. *Biochem. J.* **178,** 689-697 (1979).
8. Carrington, M. *et al. J. molec. Biol.* **221,** 823-835 (1991).
9. Rice-Ficht, A. C., Chen, K. K. & Donelson, J. E. *Nature* **294,** 53-57 (1981).
10. Miller, E. N., Allan, L. M. & Turner, M. J. *Molec. biochem. Parasit.* **13,** 309-322 (1984).
11. Cross, G. A. M. *J. cell. Biochem.* **24,** 79-90 (1984).
12. Freymann, D. M. thesis, Harvard Univ. (1988).
13. Freymann, D., *et al. J. molec. Biol.* **216,** 141-160 (1990).
14. Blum, M. L. thesis, Harvard Univ. (1990).
15. Bricogne, G. *Acta crystallogr.* **A32,** 832-847 (1976).
16. Metcalf, P., Down, J. A., Turner, M. J. & Wiley, D. C. *J. biol. Chem.* **263,** 17030-17033 (1988).
17. Rossmann, M. G. & Argos, P. *J. molec. Biol.* **105,** 75-95 (1976).
18. Chothia, C. & Lesk, A. M. *EMBO J.* **5,** 823-826 (1986).
19. Olafson, R. W. *et al. Molec. biochem. Parasit.* **12,** 287-298 (1984).
20. Lalor, T. M. *et al. Proc. natn. Acad. Sci. U.S.A.* **81,** 998-1002 (1984).
21. Strickler, J. E. *et al. Biochemistry* **26,** 796-805 (1987).
22. Reinitz, D. M., Aizenstein, B. D. & Mansfield, J. M. *Molec. biochem. Parasit.* **51,** 119-132 (1992).
23. Lesk, A. M., Levitt, M. & Chothia, C. *Prot. Engng* **1,** 77-78 (1986).
24. Needleman, S. B. & Wunsch, C. D. *J. molec. Biol.* **48,** 443-453 (1970).
25. McLachlan, A. D. *J. molec. Biol.* **61,** 409-424 (1971).
26. Cohen, C. *et al. Nature* **311,** 169-171 (1984).
27. Allen, G. & Gurnett, L. P. *Biochem. J.* **209,** 481-487 (1983).
28. Hublart, M., Mendonça-Previato, L., Boutignon, F., Huet-Duvillier, G. & Degand, P. *Comp. Biochem. Physiol.* **92B,** 705-710 (1989).
29. Wiley, D. C. & Skehel, J. J. *A. Rev. Biochem.* **56,** 365-394 (1987).
30. Rossmann, M. G. *et al. Nature* **317,** 145-153 (1985).
31. Baltz, T. *et al. EMBO J.* **10,** 1653-1659 (1991).
32. Borst, P. & Greaves, D. R. *Science* **235,** 658-667 (1987).
33. Frasch, A. C. C., Borst, P. & Van den Burg, J. *Gene* **17,** 197-211 (1982).
34. Guyaux, M., Cornelissen, A. W. C. A., Pays, E., Steinert, M. & Borst, P. *EMBO J.* **4,** 995-998 (1985).
35. Roost, H.-P., Charan, S. & Zinkernagel, R. M. *Eur. J. Immun.* **20,** 2547-2554 (1990)
36. Phillips, E. P. *et al. Nature* **354,** 453-459 (1991).
37. Reinitz, D. M. & Mansfield, J. M. *Infect. Immun.* **58,** 2337-2342 (1990).
38. Lesk, A. M. & Chothia, C. *J. molec. Biol.* **160,** 325-342 (1982).
39. Donelson, J. E. & Rice-Ficht, A. C. *Microbiol. Rev.* **49,** 107-125 (1985).
40. Orcutt, B. C., Dayhoff, M. O., George, D. G. & Barker, W. C. *ALIGN* (Protein Identification Resource, National Biomedical Research Foundation, Georgetown University Medical Center, Washington DC, 1986).
41. Reddy, L. V., Hall, T. & Donelson, J. E. *Biochem. biophys. Res. Commun.* **169,** 730-736 (1990).
42. Pays, E. *et al. Cell* **34,** 371-381 (1983).
43. Lenardo, M. J., Rice-Ficht, A. C. Kelly, G., Esser, K. M. & Donelson, J. E. *Proc. natn. Acad. Sci. U.S.A.* **81,** 6642-6646 (1984).
44. Boothroyd, J. C., Paynter, C. A., Coleman, S. L. & Cross, G. A. M. *J. molec. Biol.* **157,** 547-556 (1982).
45. Thon, G., Baltz, T. & Eisen, H. *Genes Dev* **3,** 1247-1254 (1989).
46. Rice-Ficht, A. C., Chen, K. K. & Donelson, J. E. *Nature* **298,** 676-679 (1982).
47. Metcalf, P., Blum, M., Freymann, D., Turner, M. & Wiley, D. C. *Nature* **325,** 84-86 (1987).
48. Jones, T. A. *Meth. Enzym.* **115,** 157-171 (1985).
49. Richardson, J. S. *Adv. Prot. Chem.* **34,** 167-339 (1981).
50. Priestle, J. P. *J. appl. Crystallogr.* **21,** 572-576 (1988).
51. Subbiah, S. & Harrison, S. C. *J. molec. Biol.* **209,** 539-548 (1989).
52. Devereux, J., Haeberli, P. & Smithies, O. *Nucleic Acids Res.* **12,** 387-395 (1984).
53. Kabsch, W. & Sander, C. *Biopolymers* **22,** 2577-2637 (1983).

ACKNOWLEDGEMENTS. We acknowledge the earlier contributions of P. Metcalf and D. Freymann. This work was supported by the NIH (D.C.W.). J.A.D. was supported by a Fellowship from Merck, Sharpe & Dohme. D.C.W. is an investigator of the Howard Hughes Medical Institute.