Comparative sequence analysis and predictions for the envelope glycoproteins of foamy viruses

George Wang\(^2\) and Mark J. Mulligan\(^1,2\)

Departments of Medicine\(^1\) and Microbiology\(^2\), University of Alabama at Birmingham, 845 19th Street South, BBRB 220, Birmingham, AL 35294-2170, USA

The foamy viruses (FVs) are a genus of complex retroviruses that has recently been found to possess several novel molecular features. There is increasing interest in the development of FVs as novel vectors for gene delivery. As there are remarkably few published studies of FV proteins, these recent findings prompted us to predict the structural features of FV glycoproteins with the aid of computer programs. We analysed all seven available FV Env sequences, a greater number of sequences than in previously published analyses. The relative rates of change for FV structural proteins were Pol $<$ Env $<$ Gag in increasing order, which differs from all other retroviruses. We determined that this difference is primarily caused by a higher relative rate of change for FV Gag proteins. We analysed the functional domains of FV glycoproteins and found that their structural organization was generally similar to other retroviruses. Putative structures were identified for the signal peptide, cleavage site, fusion peptide, membrane-spanning domain and the unique endoplasmic reticulum retrieval signal. Based on the predicted secondary structure of the transmembrane glycoprotein (TM) subunit, gp47, we also identified a unique prolonged central ‘sheets and loops’ region as the dominant feature of an unusually lengthy TM ectodomain. This lengthy central domain was flanked at each end by \(\alpha\)-helices. The predictions reported here will stimulate and facilitate experimental approaches to better understand the structure and function of FV glycoproteins, and should assist in the planning and development of FV vectors.

Introduction

Foamy viruses (FV) are ubiquitous in bovines, felines and non-human primates, but are not associated with any diseases (Loh, 1993; Schweizer et al., 1995). Zoonotic infections of humans exposed to primates occur without any recognized clinical sequelae (Schweizer et al., 1995, 1997; Goepfert et al., 1996; Neumann-Haefelin & Schweizer, 1997; Weiss, 1998; Heneine et al., 1998). While they are the least studied genus of retroviruses, FVs have recently become the subject of increasing interest owing to several unique biological features and for their potential use as vectors for gene replacement therapy. For example, FV Pol proteins are translated from a spliced mRNA instead of as Gag–Pol precursors (Esslle et al., 1996; Löchelt & Flügel, 1996; Yu et al., 1996); nascent Gag proteins are transiently transported into the nucleus (Schliephake & Rethwilm, 1994); and the functional nucleic acid of FV extracellular particles is linear double-stranded DNA (Moebes et al., 1997). Relatively little has been published on the structure or functions of FV glycoproteins. We recently reported that FV glycoproteins possess an endoplasmic reticulum (ER) retrieval signal (Goepfert et al., 1995) that localized the human FV (HFV) glycoprotein to the ER (Goepfert et al., 1997), a novel finding relative to all other retroviral glycoproteins. In the present study, we used computer-based methods (Livingstone & Barton, 1996; Rost & Sander, 1993; Russell & Sternberg, 1995; Gatot et al., 1998) to perform updated sequence analyses and predictions based on the seven available sequences – a greater number than has previously been analysed – of FV glycoproteins in comparison to glycoproteins from other virus families. The results obtained in this work will stimulate and guide future molecular studies of FV glycoproteins.

Methods

The protein sequences were obtained directly from SWISS-PROT or translated from the GenBank database. For the Env proteins of HFV (Flügel et al., 1987) and simian FV type 1 (SFV-1) (Mergia et al., 1990), we used the first translational start codon (Kupiec et al., 1991) instead of the
second for two reasons: (1) the first ATG had a better Kozak context than the second (Kozak, 1989), and (2) the translated protein sequences were highly conserved.

Sequence alignments and comparisons were made with programs in the Genetics Computer Group (GCG) package (Genetics Computer Group, 1997). Numbering of the sequences referred to the aligned amino acid sequence of FV Env including the signal peptide. Helical net (Lim, 1978) and wheel (Schiffer & Edmundson, 1967) analyses were made by using the Helmet and Wheel programs (Jones et al., 1992) run on VAX computers. Hydrophobicity indices (H.I.) were calculated with the normalized consensus scale of Eisenberg (1984). The signal peptide cleavage site prediction was made using SignalP, available on the Internet at http://www.cbs.dtu.dk/services/SignalP/ (Nielsen et al., 1997). The membrane-spanning domain, protein secondary structure and solvent accessibility predictions were made by using Predicis, PHDsec, PHDacc and PHDhtm – available on the Internet at http://www.embl-heidelberg.de/predicis/predictprotein/predictprotein.html (Rost & Sander, 1994 a, b; Rost et al., 1995; Rost, 1996). Hydrophobic cluster analysis (HCA) is available on the Internet at http://www.lmcp.jussieu.fr/~soyer/www-hca/hca-form.html (Gaboriaud et al., 1987; Lemesle-Varloot et al., 1990; Callebaut et al., 1997). The sequence similarity search was performed using BLAST 2, available on the Internet at http://www.ncbi.nlm.nih.gov/BLAST (Altschul et al., 1997).

**Results**

**FV Gag proteins are more variable than FV Env glycoproteins**

We began our analysis by searching the NCBI nr database (Non-redundant GenBank CDS translations + PDB + SWISS-PROT + PIR) with BLAST 2 (Altschul et al., 1997) and the SWISS-PROT database with FastA (Pearson & Lipman, 1990) for proteins homologous to the HFV glycoprotein (Flügel et al., 1987; Maurer et al., 1988). As expected, all the hits with statistically significant scores were FV glycoprotein sequences (the highest E value (least significant) was 2 × 10⁻²⁶, which was a partial SFV-1 Env protein, with identity at 59 of 93 residues).

The FV glycoproteins identified were from SFV of chimpanzees (SFVcpz) (Herchenröder et al., 1994), SFV-1 (Mergia et al., 1990), SFV type 3 (SFV-3) (Renne et al., 1992), feline FV (FeFV) (Winkler et al., 1997), feline syncytial virus (FeSV) (C. R. Helps & D. A. Harbour, accession no. U85043) and bovine syncytial virus (BSV) (Renshaw & Casey, 1994; Holzschu et al., 1998). Among these seven protein sequences, the highest percentage identity was 90-1 (between HFV and SFVcpz), the lowest was 40-4 (between SFV-3 and FeFV), and the average was 52-9 ± 16-0. We created a phylogenetic tree (Fig. 1), which prompted us to also perform searches for proteins homologous to either FeFV or BSV glycoproteins which had greater evolutionary distance from HFV. Similar results were obtained (not shown). This suggested that FVs are highly divergent from other virus families. Previous analyses of Pol protein sequences had suggested relatedness between FV and murine leukaemia virus (MuLV) (Lewe & Flügel, 1990; Renne et al., 1992; Dias et al., 1996), but this was not observed with the glycoprotein sequences (Pearson, 1996).

**Table 1. Relative rates of change of Gag and Env proteins**

Relative rate of change is defined as the ratio of the evolutionary distance (calculated amino acid substitutions per 100 amino acids) for the indicated protein to that of the Pol protein. Each result was an average of 8 to 22 determinations, presented with standard deviation. Evolutionary distances were generated by the Distances program of the SWISS-PROT package, following alignment by Pileup and arbitrary trimming of gapped regions. The Distances program used the formula of Kimura (1983): distance = −ln(1 − D − 0.2D²), where S = exact matches/positions scored, and D = 1 - S.

<table>
<thead>
<tr>
<th></th>
<th>FV*</th>
<th>MuLV</th>
<th>PIV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pol</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Gag</td>
<td>2.3 ± 0.2</td>
<td>1 ± 0</td>
<td>1.2 ± 0.3</td>
</tr>
<tr>
<td>MA</td>
<td>2.5 ± 0.5</td>
<td>2 ± 0.2</td>
<td>1.3 ± 0.4</td>
</tr>
<tr>
<td>CA</td>
<td>1 ± 0.2</td>
<td>0 ± 0.1</td>
<td>0.8 ± 0.2</td>
</tr>
<tr>
<td>NC</td>
<td>3 ± 0.7</td>
<td>0.7 ± 0.2</td>
<td>1.3 ± 0.3</td>
</tr>
<tr>
<td>Env</td>
<td>1 ± 0.0</td>
<td>2 ± 0.1</td>
<td>2 ± 0.7</td>
</tr>
</tbody>
</table>


A general rule for retroviruses is that their Pol, Gag and Env proteins have increasing rates of change in that order (McClure et al., 1988). The results we obtained for MuLV and primate immunodeficiency viruses (PIVs) agreed with this rank order (Pol < Gag < Env) (Table 1). However, the relative rates of change for FVs were Pol < Env < Gag. The rate of change for Env proteins of FVs was 1.8, which was compatible with 2.0
Foamy virus glycoproteins

Fig. 2. FV glycoproteins possess the general features of a retroviral glycoprotein. (A) The glycoprotein of the prototypic HFV was used for the hydrophobicity plot, which was generated by Pepplot of the GCG package. Both Kyte–Doolittle (window size = 9 residues) and Goldman (window size = 20 residues) methods were used. (B) A similarity plot of the seven aligned primate FV Env protein sequences was generated by the Psortsimilarity program of the GCG package. (C) Schematic diagram of the overall structure of the FV glycoprotein precursor (gp130). Residues in the hydrophobic domain of the signal peptide, the RKK/RPR cleavage signal, the conserved hydrophobic residues in the potential fusion peptide, the charged residues in the membrane-spanning domain, and the unique, conserved ER retrieval signal at the glycoprotein C terminus are in bold. Arrows indicate the predicted cleavage sites for the signal peptide and the SU/TM junction. The fusion peptide shows a typical $i, i+3/4, i+7$ pattern which is characteristic of amphipathic $\alpha$-helicies (Livingstone & Barton, 1996). The conserved residues and predicted helical content of the membrane-spanning domain are shown. Abbreviations: PHDhtm, helix transmembrane domain prediction; H, predicted helical region; Relhtm, estimated reliability of the helical prediction on a 0 to 9 scale.
for MuLV and PIVs. However, the rate of change for Gag proteins of FVs was 2–3, much higher than the 1–6 for MuLV and 1–2 for PIV. If one assumes that the rates of change for all the retroviral Pol proteins were constant, then one can infer that it was the high variability of the Gag of FVs that made the rank order of change in FV structural proteins different from that of other retroviruses. Further analysis of the relative rates of change within the functional domains of the Gag proteins revealed that the MA of FV was comparable to MuLV, with both being higher than the MA of PIVs (Table 1). However, the rates of change in both CA and NC of FVs were much higher than those of either MuLV or PIV. The relative rates of change when normalized to the corresponding reverse transcriptases (RT) (data not shown) were similar to the results obtained here with normalization to the entire Pol, but had larger standard deviations in general.

FV Env glycoproteins possess the general features of retroviral glycoproteins

As seen in the hydrophobicity plot, two prominent hydrophobic regions were identified (Fig. 2a). Analogous to other retroviral Env glycoproteins, the first of these was the hydrophobic domain of the signal peptide and the second the membrane-spanning domain (Figs 2c and 4a). A conserved RXK/RR subtilisin-like protease cleavage site was identified, in a position appropriate to divide the glycoprotein into its mature surface (SU; gp80) and transmembrane (TM; gp47) subunits (Figs 2c and 4a). We recently performed site-directed mutagenesis of this site within the HFV env which confirmed that this sequence is the FV glycoprotein cleavage site (K. L. Shaw, A. Bansal & M. J. Mulligan, unpublished results). The signal peptide was predicted to be cleaved between cysteine and phenylalanine by the SignalP program (Fig. 2c) (Nielsen et al., 1997), which conformed to the –1 –3 (small and neutral) rule (Nothwehr & Gordon, 1990; Perlman & Halvorson, 1983; von Heijne, 1986).

Potential fusion peptide. A prominent feature of the TM subunits of retroviral Env glycoproteins is the fusion peptide (Gallaher, 1987). We identified a highly conserved segment downstream from the RXK/RR cleavage signal as a putative fusion peptide based on the criteria of White (1990, 1992) (Fig. 2f). This putative fusion peptide was close to the TM protein N terminus, although not immediately N-terminal as for primate lentiviruses (Fig. 2c). This region possessed a typical i,
was the lowest among the viral fusion peptides we analysed. Interestingly, unlike other retroviruses, the FV fusion peptide did not appear as a prominent hydrophobic region in the hydrophobicity plot (Fig. 2a). This was due to its low overall average hydrophobicity per residue, or hydrophobicity index (H.I.), which was the lowest among the viral fusion peptides we analysed (Fig. 3c). However, in α-helix modelling, the fusion peptides of FVs have maintained a high H.I. at the hydrophobic face of the helix, with a score comparable to other viral fusion peptides (Fig. 3c).

Membrane-spanning domain (MSD). The second prominent hydrophobic domain near the C terminus was predicted to be the MSD by PHDhtm (Rost & Sander, 1995) (Fig. 2c). This region is similar to the MSDs of HIV and SIV, which are punctuated by positively charged residues both within and immediately C-terminal to their proposed MSDs. Using site-directed mutagenesis of HFV env, we recently confirmed experimentally that this putative MSD is necessary and sufficient for anchorage of the FV glycoprotein in cellular lipid bilayers (G. Wang & M. J. Mulligan, unpublished results). Taken together, these results indicated the topology of the FV glycoprotein is that of a type 1 membrane protein.

The unique central ‘sheets and loops’ region of FV TM proteins

The predictions for secondary structure and solvent accessibility for the TM (gp47) proteins of FVs were made with the Profile Neural Network System (Rost, 1996; Rost & Sander, 1994a, b), and were based on multiple alignment of the seven FV glycoprotein sequences made by the Pileup program of GCG. These predictions revealed a division of three structural regions within TM between the fusion peptide and the MSD: a long N-terminal α-helix region immediately downstream from the fusion peptide, a second C-terminal α-helix region just upstream from the MSD, and a lengthy central region which consisted of β-sheets and loops (Fig. 4a, b). A similar N-terminal α-helix domain was previously proposed to form the structural backbone of viral fusion glycoproteins (Bullough et al., 1994; Chambers et al., 1990; Gallaher et al., 1989; Wilson et al., 1981) whereas peptide homologues to the second α-helix domain of HIV showed a potent inhibitory effect on membrane fusion induced by the HIV-1 glycoprotein (Matthews et al., 1994). These two predicted α-helix domains in FVs appeared in position and length to the two α-helices observed in the solved structures of influenza virus HA2 and HIV-1 gp41 (Fig. 4b).

However, the long central ‘sheets and loops’ domain was unique and distinguished the TM proteins of FVs from those of all other retroviruses. This unusual long central region increased the length of the extracellular domain of the FV TM to more than twice the length of that of any other retrovirus (Fig. 5).

Discussion

Recent reports have described distinctive biological features of FVs which distinguish this retroviral genus from all others (Ensset al., 1996; Goepfert et al., 1995, 1997; Löchelt & Flügel, 1996; Schliephake & Rethwilm, 1994; Yu et al., 1996; Moebes et al., 1997). We found that the relative rates of change for FV structural proteins were Pol < Env < Gag. That result stands in contrast to Pol < Gag < Env observed with other retroviruses, e.g. for MuLV and PIV. Our results were based on analyses of seven FV sequences, and agreed with an earlier observation based on three FV sequences (Renne et al., 1992). This inversion of retroviral Env and Gag conservation could be caused either by a high conservation of FV Env, a low conservation of FV Gag, or both. Our results suggested that it was primarily caused by the low conservation of FV Gag. This conclusion was based on the rate of change calculations which assumed that retroviral Pol proteins were relatively constant compared to Gag or Env, and allowed a comparison of Env and Gag proteins with their rate of change normalized to their Pol proteins. The retroviral enzymes encoded by Pol play a vital role throughout the replication cycle of the retroviruses, which imposes the most stringent functional and structural constraints on Pol. And indeed they were found to be the most conserved structural proteins of the retroviruses, and no obvious recombinational events were observed (McCure et al., 1988). The cause of the low conservation (or the high variability) of the FV Gag proteins is unknown. We speculated that this could be related to the separation of Gag and Pol translation in FVs, as opposed to translation of a Gag–Pol polyprotein in all the other retroviruses. In the latter case, co-translation of the structurally stringent Pol protein with Gag may impose constraints on Gag variation within the Gag–Pol polyprotein. Mutations within Gag might affect Pol and/or Gag–Pol functions in the Gag–Pol polyprotein as was reported for HIV-1 (Huang et al., 1997).

Despite their sequence divergence from other retroviruses, the FV glycoproteins were found to possess the general structural features observed in all retroviral glycoproteins. Another research group recently aligned five FV glycoproteins and made a similar observation (Holzschu et al., 1998). The putative FV fusion peptide is slightly internal relative to the glycoprotein cleavage site, similar to that of Rous sarcoma virus but different from most retroviruses. The FV fusion peptide had a high propensity to form α-helical structures. The amphipathicity of the fusion peptide α-helix was shown by the high hydrophobic moment as determined by the helical wheel analysis. Despite its low overall H.I. compared to other viral fusion peptides, the fusion peptide of FVs has maintained a high H.I. at the hydrophobic face in α-helix modelling. Among the fusion peptide parameters we analysed for the five viruses,
Fig. 4. (A) For legend see facing page.
Fig. 4. (A) Prediction for secondary structure and solvent accessibility of the seven available FV TM glycoproteins. The predicted RXK/RR cleavage site between SU and TM is shown for clarity (bold). Abbreviations: AA#, amino acids numbered according to the aligned FV glycoprotein sequences; PhDsec, secondary structure predictions (H = helix, E = extended sheet, blank = loop); P 3 acc, predicted relative solvent accessibility (e = exposed, b = buried, blank = intermediate). (B) Schematic diagrams of the HA2 glycoprotein of influenza A virus and the TM glycoproteins of HIV-1 and FV. Empty boxes represent fusion peptides; hatched boxes, α-helices; solid boxes, membrane-spanning domains. Horizontal lines between hatched boxes represent ‘sheets and loops’ regions. Conserved cysteines (c) and conserved potential N-linked glycosylation sites (†) are shown. For simplicity, cytoplasmic domains are not shown. Predictions were made with the Profile Neural Network System (Rost & Sander, 1993, 1994a, b). Twenty protein sequences were used to generate the diagram for influenza virus HA2; five were used for HIV-1; and seven were used for FV. The secondary structure predictions for influenza virus HA2 and for the HIV TM were consistent with their structures solved by crystallography (Bullough et al., 1994; Chan et al., 1997; Weissenhorn et al., 1997). Similar secondary structure predictions for FV TM were obtained with a second technique, hydrophobic cluster analysis (HCA) (data not shown) (Gatot et al., 1998). It is worth noticing that potential disulfide bonds exist as a result of the seven conserved cysteines in the central ‘sheets & loops’ region of the FV TM.

Fig. 5. Comparison of the glycoproteins of several retroviruses and influenza A virus. Horizontal lines and boxes represent mature polypeptide chains (signal peptides not shown) with their lengths drawn to scale. SU proteins are shown to the left of the cleavage site (arrow), TM proteins to the right. Empty boxes represent fusion peptides; solid boxes represent membrane-spanning domains (MSD). The TM cytoplasmic domains are not drawn, but their lengths are indicated after the ‘–’ to the right of the solid boxes (MSD). The length of the TM protein extracellular domain is indicated in the column on the right. The HFV TM extracellular domain (376 aa) is more than twice the length of the others.
the high H.I. at the hydrophobic face was the only consensus we observed. It was reported that the hydrophobicity gradient and the oblique orientation of the fusion peptides were critical for their fusogenic activity (Brasseur et al., 1988, 1990). Mutagenesis that altered the theoretical angle of the fusion peptides decreased the fusion ability (Vonieche et al., 1992). Interestingly, fusion peptides had an unusual content of small amino acids, which may contribute to their conformational mobility to allow them to adopt multiple secondary structures (Callebaut et al., 1997; Durell et al., 1997). Mutagenesis studies of the FV putative fusion peptide are under way in our laboratory.

It is worth noticing that the ER retrieval signal (lysine at −3, lysine or arginine at −4 or/and −5 relative to the C terminus) is conserved for all seven FV glycoproteins (Jackson et al., 1990; Shin et al., 1991; Goepfert et al., 1995). Studies of the significance of the ER retrieval signal for the FVs are under way in our laboratory (Goepfert et al., 1997).

Based on the predicted secondary structure of the FV TM protein (gp47), we identified an unusually prolonged central ‘sheets and loops’ region that distinguished the FV TM from that of all other retroviruses. The resulting lengthy TM protein extracellular domain must be responsible for the distinctive appearance of FVs under the electron microscope where the glycoproteins appear as very prominent, regularly spaced, long spikes. The biological function provided to FVs by this distinctive glycoprotein structure is not known.

All retroviral RTs are highly conserved and therefore were assumed to share a common evolutionary ancestry. Sequence divergence among the less conserved Env glycoproteins of divergent retrovirus families has occurred presumably due to antigenic escape driven by their hosts’ immune systems or due to adaptation to the host cellular receptor. Despite this sequence divergence, all retroviral TM glycoproteins share certain conserved features suggesting a common structural and functional organization (Coffin, 1986; Hunter & Swanstrom, 1990). The recently solved structures of the TM glycoproteins of MuLV (Fass & Kim, 1996) and HIV (Chan et al., 1997; Weissenhorn et al., 1997) revealed striking similarity to the previously solved influenza HA2 subunit (Bullough et al., 1994). It now seems likely that the TM proteins of all retroviruses share a similar three-dimensional core structure. This TM core structure is a trimer comprising N-terminal and C-terminal α-helices, similar to the α-helices we predicted for the TMs of FVs. Mutational analyses of specific domains within the unique TM proteins of FVs, particularly the lengthy, central ‘sheets and loops’ region and the flanking N- and C-terminal α-helices, will provide new insights into the structural features of retroviral glycoproteins.

We wish to thank Drs William R. Pearson (University of Virginia) and Elliot Lefkowitz for their advice. The GCG programs package was made available by the UAB AIDS Center which is supported by U.S. Public Health Service grant AI37784 and AI28147; and Cystic Fibrosis Foundation Research Development Grant R464.

References


Received 29 June 1998; Accepted 31 August 1998