JOURNAL OF VIROLOGY, Nov. 1994, p. 7566–7569
0022-538X/94/$04.00+0
Copyright © 1994, American Society for Microbiology
Vol. 68, No. 11

Insertion of N-Linked Glycosylation Sites in the Variable Regions of the Human Immunodeficiency Virus Type 1 Surface Glycoprotein through AAT Triplet Reiteration

MARNIX L. BOSCH,[1]* ARNO C. ANDEWEG,[1]† RONALD SCHIPPER,[2] AND MARCEL KENTER[1]‡

Laboratory of Immunobiology, National Institute for Public Health and Environmental Protection, 3720 BA Bilthoven,[1] and Laboratory of Immunohaematology and Blood Bank, University Hospital E3-Q, 2333 AA Leiden,[2] The Netherlands

Received 20 May 1994/Accepted 26 July 1994

Variable regions with sequence length variation in the human immunodeficiency virus type 1 envelope exhibit an unusual pattern of codon usage with AAT, ACT, and AGT together composing >70% of all codons used. We postulate that this distribution is caused by insertion of AAT triplets followed by point mutations and selection. Accumulation of the encoded amino acids (asparagine, serine, and threonine) leads to the creation of new N-linked glycosylation sites, which helps the virus to escape from the immune pressure exerted by virus-neutralizing antibodies.

The surface glycoprotein gp120 of the human immunodeficiency virus type 1 (HIV-1) contains a number of regions with a relatively high degree of amino acid sequence variation, designated variable regions 1 through 5 (V1 to V5 [11]). Recently, evidence has been presented that, like variation in, e.g., the V3 region, variation in the V1 and V2 regions contributes to the determination of viral tropism and cytopathic potential (2, 3, 8, 17, 19), indicating that these regions are probably involved in the process of envelope-mediated membrane fusion resulting in virus entry or in syncytium formation. Sequence variation in these regions can then be reflected in these processes, resulting in variations in cell tropism and cytopathicity as described above. Virus-neutralizing antibodies to the V2 region that probably directly interfere with virus entry have also been described previously (7, 10a, 14), and sequence variation could enable the virus to escape from such antibodies. The addition of carbohydrates can alter the recognition of envelope glycoproteins by the immune system (1, 6). Here we describe a process of sequence length variation in the HIV-1 envelope gene that results in the selective insertion of new N-linked glycosylation sites in the variable regions of the envelope glycoproteins as a mechanism to escape immune pressure exerted by neutralizing antibodies.

The alignments of the variable-length sequences found in V1 and V2 are shown in Fig. 1. We have used the HIV-1 envelope sequence alignment presented by Myers et al. (12) as a starting point and readjusted the alignment by eye. The numbers at the top indicate the codons incorporated in the regions of variable sequence length. The underlined blocks of three codons indicate encoded potential N-linked glycosylation sites (see below). We have noticed an apparent skewed usage of codons in these regions. A relatively high number of codons begin with A and end with T, with an especially high frequency of AAT codons.

To confirm this observation, we have counted the frequencies of the four nucleotides adenosine (A), cytidine (C), guanosine (G), and thymidine (T), at the first, second, and third positions of each codon in the variable-length sequences of the V1 and V2 regions. We find sharp deviations from the distributions as found in the whole of gp120 (we have used the HXB2R gp120 sequence for comparison). Most notably, we find that approximately 84% of the codons in these regions start with A (<1% start with C); only 1 to 4% of codons have a T at position 2 (compared with 26% in gp120); and 67 to 76% of codons end with T, with sharp decreased frequencies of A and G at position 3. We have subsequently analyzed the frequency distributions of all four nucleotides at each codon position in the whole envelope gene alignment by a sliding window method (see legend to Fig. 2 for details). The results obtained for A at position 1 (A1), T at position 2 (T2), and T at position 3 (T3) are plotted in Fig. 2 for a window size of 20 codons shifted by 1 codon, as are the number of sequences counted in each window to map the regions with sequence length variation (bottom panel). It can be seen that sharp increases in frequencies of A1 and T3 with concomitant decreases in T2 occur in the V1 and V2 regions and also in two other regions that display a high degree of sequence length variation, the V4 and V5 regions. No other regions of the envelope gene display similar concomitant nucleotide frequency distributions. The A1 and T3 peaks are not as pronounced for V5 because of the relatively small size of the region of variable sequence length in V5 relative to the window size.

We postulate here that the skewed distribution of nucleotides at the first, second, and third positions of the codons that make up the regions of variable sequence length is the result of the insertion of AAT triplets at these sites followed by point mutations and selection for sequences that provide a selective advantage to the virus by the insertion of N-linked glycosylation sites, resulting in escape from the immune system, as well as selection against disadvantageous (in this case, hydrophobic) sequences. The implications of this mechanism will be discussed below.

The use of a severely limited set of nucleotides at position 1

* Corresponding author. Present address: Regional Primate Research Center and Department of Pathobiology SC-38, University of Washington, Seattle, WA 98195. Phone: (206) 543-7619. Fax: (206) 543-3873.

† Present address: Laboratory of Viral Pathogenesis, Biomedical Primate Research Center, Rijswijk, The Netherlands.

‡ Present address: Institute of Virology, Erasmus University, Rotterdam, The Netherlands.

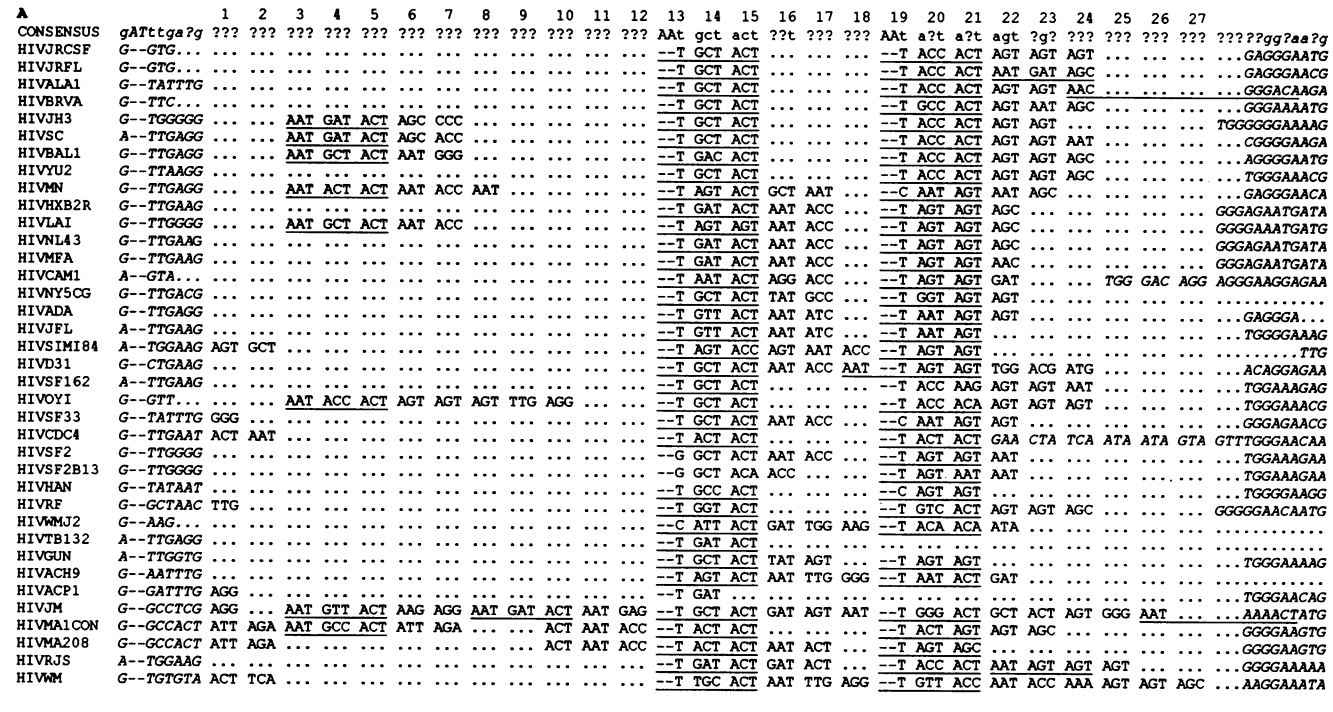

FIG. 1. Sequence alignments of variable-length sequences in V1 (A) and V2 (B). Sequences and alignment were adopted from the work of Myers et al. (12) and adjusted by hand. Numbers in the top row indicate codons of the variable-length sequence region. Sequences outside of the variable-length sequences discussed here are in italics. Dashes indicate nucleotides identical to the consensus sequence (top row), and dots indicate deletions relative to the consensus sequence. Uppercase letters in the consensus sequence indicate bases conserved throughout all aligned sequences; lowercase letters indicate bases found in at least 50% (but less than 100%) of the aligned sequences in reference 12. Blocks of three codons encoding a potential N-linked glycosylation site (asparagine-X-serine/threonine) are underlined.

(as, in this case, mostly A) is reflected in the encoded amino acid sequence. The preferential use of codons starting with an A (>80% in the V1 and V2 variable-length sequences) results in predominant insertion of isoleucine, methionine, asparagine, serine, threonine, lysine, and arginine. The virtual absence of T's at position 2 results in the exclusion of isoleucine (Ile) and methionine (Met), the only two hydrophobic amino acids whose codon starts with an A. Apparently, the insertion

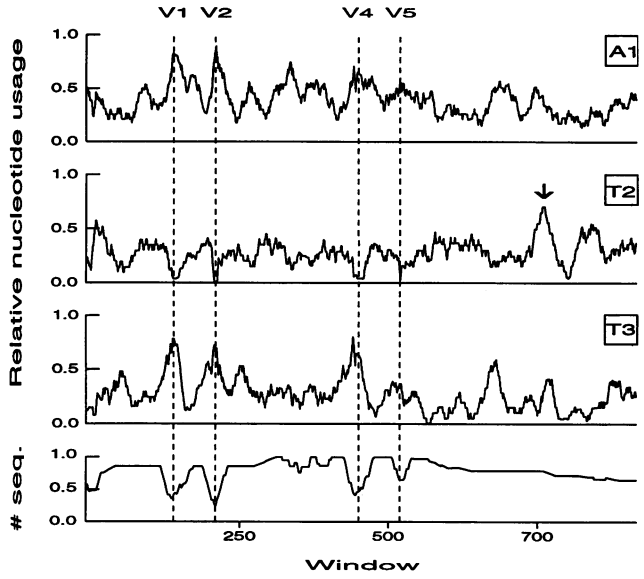

FIG. 2. Frequency distributions of A1 (top panel), T2 (second panel from top), and T3 (third panel from top) in the HIV-1 envelope gene, as well as the relative number of sequences counted in each window (bottom panel). Valleys in the bottom panel plot indicate regions of variable sequence length. Frequencies of each nucleotide at each position were counted in a window of 20 codons, shifted by 1 codon, with the aligned sequences taken from reference 12. Only A1, T2, and T3 are shown. Variable regions V1, V2, V4, and V5 are indicated (except V3). The 95% confidence interval upper and lower limits, calculated as mean ± 1.96× standard deviation following square root transformation of the data to achieve normality, are as follows: A1, 0.71 and 0.15; T2, 0.56 and 0.08; and T3, 0.63 and 0.05, respectively. The A1 and T3 peaks in V1, V2, and V4 exceed these limits, as do the T2 valleys in V1, V2, V4, and V5. A number of A1 peaks are found besides the V1, V2, V4, and V5 regions. The lack of concomitant T2 valleys and T3 peaks indicates that these have been generated through mechanisms different from the one postulated in this paper. A T2 peak is found at positions corresponding to windows 754 to 769 (arrow). All codons with a T at position 2 code for hydrophobic amino acids, and this peak corresponds to the location of the hydrophobic membrane anchor sequences in gp41. Both the alignments and the computer program used to calculate this plot are available from the authors upon request.

of hydrophobic amino acids in this region of env is disadvantageous to the virus. The strong bias towards T, a remnant of the postulated originally inserted AAT, and to a lesser extent towards C as a result of mutation and selection, at position 3 now selectively inserts threonine, asparagine, and serine at these positions. Both threonine and serine codons can be generated from the AAT and AAC asparagine codons through one mutation. Four of the six codons that can potentially code for serine start with a T rather than with an A, namely, T-C-x. These codons are not used in the insertion sequences described here, although they normally compose approximately 60% of all serine codons in mammalian genes (18) and are used throughout the Env open reading frame (not shown). Random insertions and mutations followed by selection would have resulted in a predominant use of T-C-x codons for serine. The exclusive use of AGT and AGC codons to code for serine is therefore a strong point in favor of the postulated mechanism. We have counted the frequency of asparagine, threonine, and serine in V1 and V2 and compared them with those in the gp120 amino acid sequence of HIV_{HXB2R}. In the V1 and V2 variable-length sequences, the combination of Asn, Thr, and Ser makes up 75% of all amino acids, compared with only 22% of all amino acids in gp120.

Insertion of AAT triplets followed by mutation and selection results in the insertion of predominantly asparagine (N), threonine (T), and serine (S) as described above. Random combination of these three amino acids will result in the frequent occurrence of N-x-T and N-x-S sequence blocks, where "x" stands for any amino acid. Both N-x-T and N-x-S are potential N-linked glycosylation sites, and indeed the variable-length sequences are characterized by the occurrence of multiple potential glycosylation sites. Approximately 55% of all amino acids in the variable-length sequences are actually involved in the formation of N-linked glycosylation sites (Fig. 1), demonstrating the efficiency of the postulated mechanism for the generation of such sites. If glycosylation of these regions confers selective advantage to the virus, then it might be expected that virus genotypes that contain such a novel glycosylation site would persist and appear in future virus generations. Examples of such glycosylation sites that have become stable features of the HIV-1 genome can be found in both the V1 and V2 variable-length sequences (V1 codons 13 to 15 and 19 to 21 and V2 codons 16 to 18 [Fig. 1]) and the V4 and V5 regions (not shown). The occurrence of sequences with variable length 5′ and 3′ of these conserved sites indicates that the mechanism of AAT triplet insertion is ongoing; we postulate that the accumulation of more glycosylation sites provides further selective advantages to the virus through better masking of the epitopes in this region from the immune system. We predict, therefore, that the number of glycosylation sites in these regions of the Env glycoproteins of viruses isolated from infected individuals is directly related to the immune pressure exerted on these regions, which may vary from person to person, but in all probability will increase over time, especially during the asymptomatic period of infection. Others have shown a statistically significant sequence length increase and addition of a novel N-linked glycosylation site at the V2 insertion sequence that correlates with increased viral virulence (8). More virulent virus variants generally arise over time in infected individuals and, as discussed above, so will the length of the insertion sequences because of selective advantages of the viruses that have acquired more glycosylation sites. The observed correlation between viral virulence and sequence length variation in V2, therefore, probably reflects the parallelism in time of these two processes. Accumulation of sequences rich in potential glycosylation sites (both N-linked and O-linked) has been reported for the V1 regions of simian immunodeficiency virus (SIV_{MNE}) isolates obtained from monkeys that progressed towards AIDS (13). This region in SIV shows evidence of ACA triplet reiteration (based on the SIVenv alignment in the work of Myers et al. [12] and on the work of Overbaugh and Rudensey [13], not shown), which results in the accumulation of threonine residues, thereby creating potential targets for O-linked glycosylation (13). It appears that the highly related lentiviruses SIV and HIV have adopted very similar but not identical strategies (AAT versus ACA reiteration) to achieve the same goal: glycosylation of sites on the envelope glycoproteins that are important for the virus life cycle, thereby shielding these sites from recognition by antibodies.

The sequences in the V1 and V2 region play a direct role in virus entry, as discussed above. No such functional properties have been described for either the V4 or the V5 region, and yet it appears from our analysis that they may be subject to the same selective pressures as V1 and V2. The high degree of variation in these regions hampers functional studies, e.g.,

because of the difficulties of raising cross-reactive antibodies. It is conceivable that functional roles for V4 and V5 will be described in the future. Alternatively, glycosylation at V4 and V5 could help mask other functional regions that are brought into the proximity of V4 and/or V5 in the folded molecular complex.

The mechanism by which the described triplet reiterations arise is unclear. Recently, a number of human genetic diseases characterized by trinucleotide repeats have been described (reviewed in reference 15), and strand slippage during DNA replication (5) has been proposed as a mechanism that could introduce such repeats (16). Tandemly repeated sequences are also observed in human hypervariable minisatellite sequences (10), and an alternative mechanism involving double-stranded breaks and gap repair has been proposed (9). Although it is tempting to speculate that the viral reverse transcriptase is primarily responsible for the observed phenomenon, in vitro studies of the HIV-1 reverse transcriptase have not revealed direct evidence for triplet insertion, although strand slippage during replication of DNA or RNA is observed (4). The examples quoted above are compatible with the idea that cellular polymerases may be involved.

We are greatly indebted to Kees Siebelink for helpful suggestions and critical reading of the manuscript and to Albert Osterhaus for continued support.

REFERENCES

1. **Alexander, S., and J. H. Elder.** 1984. Carbohydrate dramatically influences immune reactivity of antisera to viral glycoprotein antigens. Science **226:**1328–1330.
2. **Andeweg, A. C., P. Leeflang, A. D. M. E. Osterhaus, and M. L. Bosch.** 1993. Both the V2 and V3 regions of the human immunodeficiency virus type 1 surface glycoprotein functionally interact with other envelope regions in syncytium formation. J. Virol. **67:** 3232–3239.
3. **Boyd, M. T., G. R. Simpson, A. J. Cann, M. A. Johnson, and R. A. Weiss.** 1993. A single amino acid substitution in the V1 loop of human immunodeficiency virus type 1 gp120 alters cellular tropism. J. Virol. **67:**3649–3652.
4. **Boyer, J. C., K. Bebenek, and T. A. Kunkel.** 1992. Unequal human immunodeficiency virus type 1 reverse transcriptase error rates with RNA and DNA template. Proc. Natl. Acad. Sci. USA **89:** 6919–6923.
5. **Chamberlin, M., and P. Berg.** 1962. Deoxyribonucleic acid-directed synthesis of ribonucleic acid by an enzyme from Escherichia Coli. Proc. Natl. Acad. Sci. USA **48:**81–88.
6. **Davis, D., D. M. Stephens, C. Willers, and P. Lachmann.** 1990. Glycosylation governs the binding of antipeptide antibodies to regions of hypervariable amino acid sequence within recombinant gp120 of human immunodeficiency virus type 1. J. Gen. Virol. **71:**2889–2898.
7. **Fung, M., C. Sun, W. Gordon, R. S. Liou, T. W. Chang, W. Sun, E. Daar, and D. Ho.** 1992. Identification and characterization of a neutralization site within the second variable region of human immunodeficiency virus type 1 gp120. J. Virol. **66:**848–856.
8. **Groenink, M., R. A. M. Fouchier, S. Broersen, C. H. Baker, M. Koot, A. B. van 't Woudt, H. G. Huisman, F. Miedema, M. Tersmette, and H. Schuitemaker.** 1993. Relation of phenotype evolution of HIV-1 to envelope V2 configuration. Science **260:** 1513–1516.
9. **Jeffreys, A. J., K. Tamaki, A. MacLeod, D. G. Monckton, D. L. Neil, and J. A. L. Armour.** 1994. Complex gene conversion events in germline mutation at human minisatellites. Nat. Genet. **6:**136–145.
10. **Jeffreys, A. J., V. Wilson, and S. L. Thein.** 1985. Hypervariable 'minisatellite' regions in human DNA. Nature (London) **314:**67–74.

10a. **McKeating, J. A., C. Shotten, J. Cordell, S. Graham, P. Balfe, N. Sullivan, M. Charles, M. Page, A. Bolmstedt, S. Olofsson, S. C. Kayman, Z. Wu, A Pinter, C. Dean, J. Sodroski, and R. A. Weiss.** 1993. Characterization of neutralizing monoclonal antibodies to linear and conformation-dependent epitopes within the first and second variable domains of human immunodeficiency virus type 1 gp120. J. Virol. **67:**4932–4944.

11. **Modrow, S., B. H. Hahn, G. S. Shaw, R. C. Gallo, F. Wong-Staal, and H. Wolf.** 1987. Computer-assisted analysis of envelope protein sequences of seven human immunodeficiency virus isolates: prediction of antigenic epitopes in conserved and variable regions. J. Virol. **61:**570–578.
12. **Myers, G., B. Korber, S. Wain-Hobson, R. F. Smith, and G. N. Pavlakis (ed.).** 1993. Human retroviruses and AIDS. Los Alamos Laboratory, Los Alamos, N. Mex.
13. **Overbaugh, J., and L. M. Rudensey.** 1992. Alterations in potential sites for glycosylation predominate during evolution of the simian immunodeficiency virus envelope gene in macaques. J. Virol. **66:**5937–5948.
14. **Pinter, A., S. Kayman, Z. Wu, W. Honnen, H. Chen, J. McKeating, C. Shotten, S. Warrier, and S. Tilley.** 1993. Epitope mapping and functional characterization of neutralizing monoclonal antibodies directed against the V2 domain of HIV-1 gp120, abstr. Q443. J. Cell. Biochem. **1993**(Suppl. 17E):69.
15. **Richards, R. I., and G. R. Sutherland.** 1992. Dynamic mutations: a new class of mutations causing human disease. Cell **70:**709–712.
16. **Richards, R. I., and G. R. Sutherland.** 1994. Simple repeat DNA is not replicated simply. Nat. Genet. **6:**114–116.
17. **Sullivan, N., M. Thali, C. Furman, D. Ho, and J. Sodroski.** 1993. Effect of amino acid changes in the V1/V2 region of the human immunodeficiency virus type 1 gp120 glycoprotein on subunit association, syncytium formation, and recognition by a neutralizing antibody. J. Virol. **67:**3674–3679.
18. **Wada, K., S. Aota, R. Tsuchiya, F. Ishibashi, T. Gojobori, and T. Ikemura.** 1990. Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res. **18:**2367–2411.
19. **Westervelt, P., D. B. Trowbridge, L. G. Epstein, B. M. Blumberg, Y. Li, B. H. Hahn, G. M. Shaw, R. W. Price, and L. Ratner.** 1992. Macrophage tropism determinants of human immunodeficiency virus type 1 in vivo. J. Virol. **66:**2577–2582.