JOURNAL OF VIROLOGY, Jan. 1999, p. 751–753 Vol. 73, No. 1
0022-538X/99/$04.00+0

"Hidden" dUTPase Sequence in Human Immunodeficiency Virus Type 1 gp120

CHANTAL ABERGEL,* DAVID L. ROBERTSON, AND JEAN-MICHEL CLAVERIE

Laboratory of Structural and Genetic Information, CNRS EP-91, Marseille F-13402, France

Received 27 July 1998/Accepted 14 October 1998

A coding region homologous to the sequence for essential eukaryotic enzyme dUTPase has been identified in different genomic regions of several viral lineages. Unlike the nonprimate lentiviruses (caprine arthritis-encephalitis virus, equine infectious anemia virus, feline immunodeficiency virus, and visna virus), where dUTPase is integrated into the *pol* coding region, this enzyme has never been demonstrated to be present in the primate lentivirus genomes (human immunodeficiency virus type 1 [HIV-1], HIV-2, or the related simian immunodeficiency virus). A novel approach allowed us to identify a weak but significant sequence similarity between HIV-1 gp120 and the human dUTPase. This finding was then extended to all of the primate lentivirus lineages. Together with the recently reported fragmentary structural similarity between the V3 loop region and the *Escherichia coli* dUTPase (P. D. Kwong, R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson, Nature 393:648–659, 1998), our results strongly suggest that an ancestral dUTPase gene has evolved into the present primate lentivirus CD4 and cytokine receptor interacting region of gp120.

The role of the dUTPase protein is to produce dUMP to decrease the intracellular concentration of dUTP so that uracil cannot be misincorporated into DNA (12). This enzyme, essential in eukaryotes (4), has been acquired by multiple viral lineages (11). dUTPase sequences are highly variable, and no position is strictly conserved in an alignment (2) of sequences from mammals, yeast, plants, *Escherichia coli*, and viruses, making it impossible to derive a suitable consensus sequence to retrieve all dUTPases from sequence databases. On the other hand, dUTPases from the same subfamily (e.g., from lentiviruses) are too closely related to build a motif capable of retrieving distant relatives by using current methods.

Kwong et al. (6) have recently determined the structure of the human immunodeficiency virus type 1 (HIV-1) envelope glycoprotein gp120. Although this structure has no precedent, the authors noticed fragmentary structural similarities between the outer domain of gp120 and unspecified segments of the *E. coli* dUTPase (7). However, they were unable to find evidence of any sequence relationship to confirm this. Given that most nonprimate lentiviruses are known to encode a dUTPase, we further investigated this coincidence and were able to detect a subtle but significant sequence similarity between human dUTPase and the gp120 of primate lentiviruses.

A new approach that takes advantage of the high variability between dUTPases was used to capture some invariant properties of the dUTPase sequences. Starting from an alignment of lentivirus dUTPases (from caprine arthritis-encephalitis virus, equine infectious anemia virus, feline immunodeficiency virus, and visna virus; SwissProt accession no. P33459, P03371, P16088, Q84809, and P35956, respectively), we discarded all strictly conserved positions as noninformative and concentrated instead on the positions for which the contrast between residue variability and invariance of the hydropathy index (5) was the strongest. A regular expression motif, spanning 113 positions and allowing one gap of up to four residues, was then designed to capture these most informative (hydropathy-wise) positions:

[VCA][PAQ]<.3>[MTHFS]<.25>[TAIG]<.25>[QGN]<.2>[CMLI]
<.3>[GNST]<.2>[NASGE]<.27>[NSVIT]<.15><.?4>[FYI]

where $<.n>$ denotes a fixed spacing of n successive positions occupied by any of the 20 amino acids and $<.?n>$ denotes a variable spacing (i.e., gap) of 0 to n positions. Brackets enclose the choices of residues allowed at a given position.

This highly degenerate pattern was used to scan (9) the viral section of the GenBank database and located a putative dUTPase similarity in frame with the *env* gene product of two different HIV-1 strains (AL and Z3; GenBank accession no. U95476 and K03347, respectively). To assess the statistical significance of this finding, a synthetic database of one million sequences was generated by repetitive randomization of the gp120 sequence of the HIV-1 subtype B strain HXB2 (GenBank accession no. K03455). Three hundred forty-one occurrences of the pattern shown above were found, leading to an estimated P value of 3.4×10^{-4}.

The two identified HIV-1 *env*-encoded amino acid sequences were then aligned with all known dUTPases (2) to determine their closest relative. Interestingly, the highest similarity was obtained with the human dUTPase (GenBank accession no. P33316).

The human dUTPase sequence was then aligned with the gp120 consensus sequences (10) for HIV-1 groups O and M (subtypes A to H). In pairwise comparisons, each HIV-1 subtype had 21 to 25 identical residues to the human dUTPase sequence (Fig. 1). Despite the variability of HIV-1 sequences in this region, the combined multiple alignment exhibited 16 strictly conserved positions and spanned 140 of 145 residues of the human dUTPase (Fig. 1). Such a 16-residue consensus pattern was not found to occur in a synthetic database of 10 million randomized gp120 sequences, a finding that corresponds to a high statistical significance ($P < 10^{-7}$). Its biological relevance is confirmed by the fact that when the viral

* Corresponding author. Mailing address: Laboratory of Structural and Genetic Information, CNRS EP-91, 31 Chemin Joseph-Aiguier, Marseille F-13402, France. Phone: 33 4 91 16 45 48. Fax: 33 4 91 16 45 49. E-mail: chantal@igs.cnrs-mrs.fr.

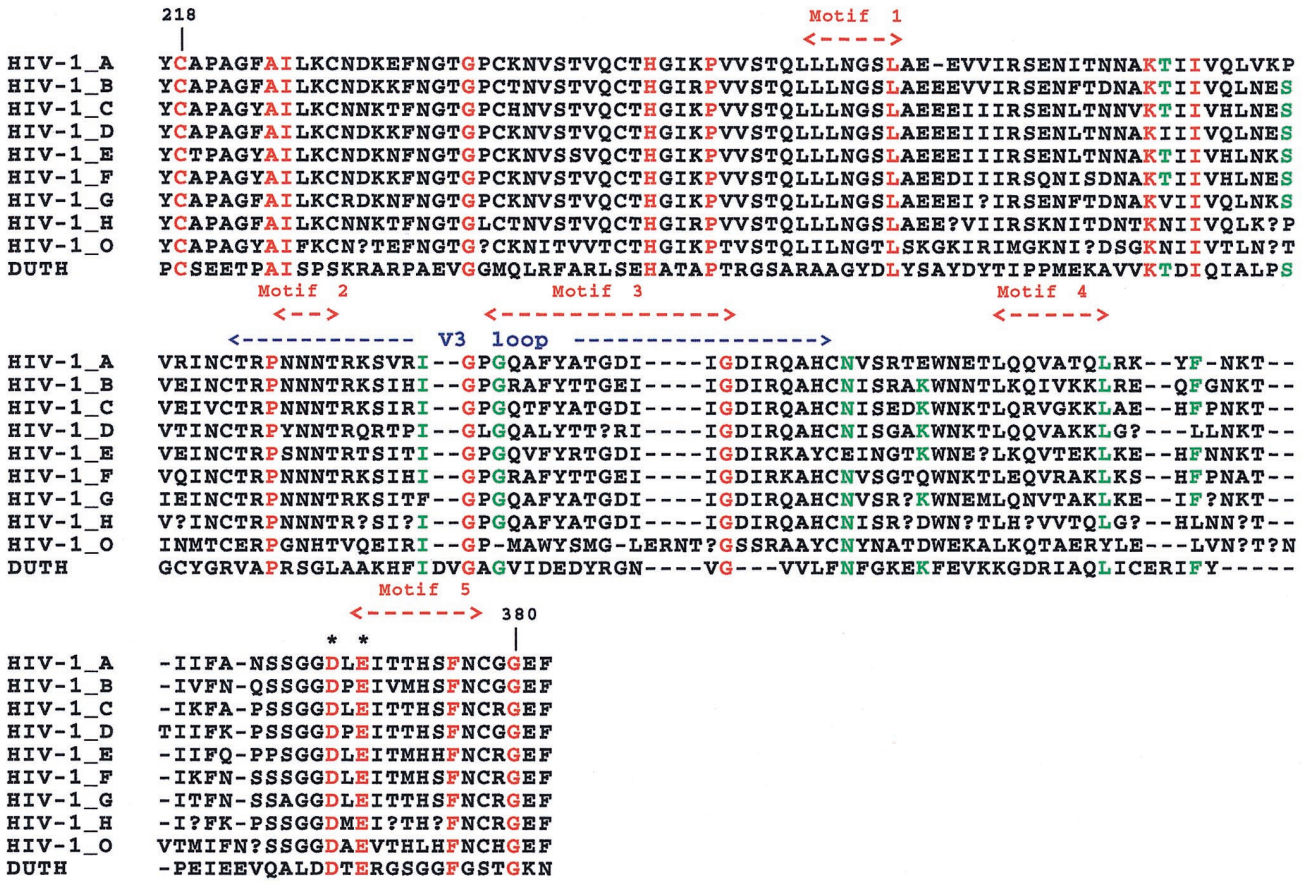

FIG. 1. Alignment of the human dUTPase (DUTH) sequence with consensus sequences (10) for HIV-1 subtypes A to H and group O. Sixteen positions shown in red are strictly conserved. Other consensus residues (over 50% in HIV-1) are shown in green. A question mark denotes the absence of a consensus residue. Numbering corresponds to that in reference 6. The V3 loop and the locations of the five previously recognized conserved motifs in dUTPases (11) are indicated above the alignment. Four of these five motifs correlate with at least one strictly conserved residue. This is also true for other primate lentivirus lineages (Fig. 2). Asterisks indicated conserved residues Asp 368 and Glu 370.

section of the GenBank database was scanned, 775 sequences (all corresponding to HIV-1 gp120) were found to exhibit this consensus.

The gp120 sequences of the various primate lentivirus lineages are very divergent, and their simultaneous alignment introduces numerous insertions and deletions, obscuring their individual relationship with the dUTPase sequence. Indeed, the pairwise alignments of human dUTPase with gp120 of various HIV-1, HIV-2, and simian immunodeficiency virus (SIV) sequences exhibit comparable levels of similarity (22 to 28 identical residues) (Fig. 2). Thus, all primate lentivirus gp120 sequences appear evolutionarily equidistant from the human dUTPase sequence.

In contrast to other lentiviruses, primate lentiviruses lack a dUTPase activity (3, 18). Other viruses are known to have incorporated the dUTPase gene at different locations (mostly in the *pol* region) in their genomes (2, 3). Thus, it is tempting to postulate that the central portions of HIV and SIV gp120 envelope proteins (residues 218 to 380 in Fig. 1) may have originated from the insertion of a mammalian dUTPase. Based on the HIV-1 gp120 three-dimensional structure (6, 14), this area involves mostly β-sheets and the V3 loop. Similarly, dUTPase structures show a predominance of β-sheets (7, 13). Nine of the strictly conserved residues (Leu 33, Lys 52, Ile 55, Pro 70, Ile 80, Gly 83, Gly 85, Gly 98, and Leu 120; numbered according to the *E. coli* dUTPase sequence [7]) cluster in the core of the dUTPase β-barrel or are located in the conserved β-sheets of the known dUTPase structures (7, 13). This suggests that their conservation is important in maintaining the overall fold. CD4 binding involves three critical residues in gp120: Asp 368, Glu 370, and Trp 427 (6). Asp 368 and Glu 370 are among those conserved in our alignment (Fig. 1) and in that with the other primate lentiviruses (Fig. 2). The V3 loop, involved in the chemokine receptor binding (14), is also remarkably central to the putatively dUTPase-derived region.

By itself, the sequence homology identified here is not sufficient to discriminate between convergent or divergent evolution. However, the fact that the human dUTPase is the closest relative to primate lentivirus gp120, and appears equidistant from these now-divergent sequences, argues that the gene encoding it was acquired by an ancestral lentivirus from its host, before the separation of the main primate lentivirus lineages. Presumably, the enzymatic role of dUTPase evolved to the current receptor function, accelerating the loss of sequence similarity while retaining part of the original fold (6). The loss of dUTPase activity may bear some evolutionary advantages for the virus, such as increasing its mutability (8, 18) and inducing its latency in nondividing infected macrophages and T cells (16, 17). Alternatively, the loss of dUTPase might have been compensated by the function of the *vpr* gene in primate lentiviruses (1).

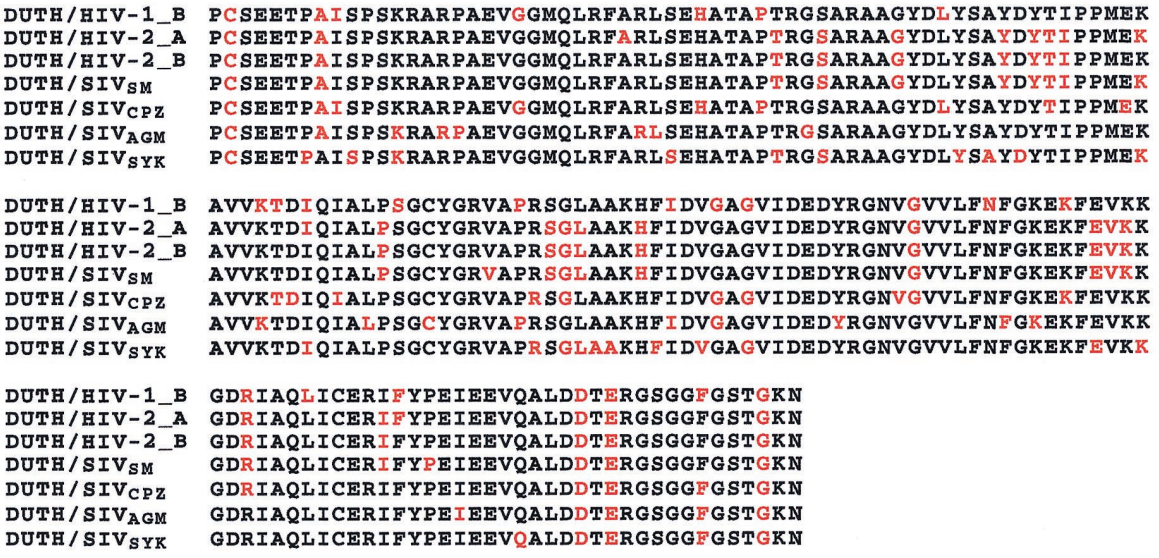

FIG. 2. Conserved positions in the pairwise alignment of human dUTPase with various HIV-1, HIV-2, and SIV gp120 sequences. The human dUTPase sequence is used as a template to visualize its relationship with gp120 sequences representative of the diversity of the primate lentiviruses (PLVs). PLV type 1 (PLV-1), HIV-1 and SIV_{CPZ}; PLV-2, HIV-2/SIV_{SM}; PLV-3, SIV_{AGM}; and PLV-5, SIV_{SYK} were used according to the nomenclature of Sharp et al. (15). Sequences for the HIV-1_B, HIV-2_A, and HIV-2_B subtype representatives are consensus sequences from the Los Alamos HIV database (10); the GenBank accession numbers for SIV_{SM}, SIV_{CPZ}, SIV_{AGM}, and SIV_{SYK} are X14307, X52154, M66437, and L06042, respectively. Conserved positions are shown in red.

We acknowledge the helpful comments of an anonymous referee.

This work was partially supported by a grant from Incyte Pharmaceuticals, Inc. D.L.R. is funded by the Agence Nationale de Recherches sur le SIDA.

REFERENCES

1. **Bouhamdan, M., S. Benichou, F. Rey, J.-M. Navarro, I. Agostini, B. Spire, J. Camonis, G. Slupphaug, R. Vigne, R. Benarous, and J. Sire.** 1996. Human immunodeficiency virus type 1 Vpr protein binds to the uracil DNA glycosylase DNA repair enzyme. J. Virol. **70:**697–704.
2. **Corpet, F., J. Gouzy, and D. Kahn.** 1998. The ProDom database of protein domain families. Nucleic Acids Res. **26:**323–326.
3. **Elder, J. H., D. L. Lerner, C. S. Hasselkuslight, D. J. Fontenot, E. Hunter, P. A. Luciw, R. C. Montelaro, and T. R. Phillips.** 1992. Distinct subsets of retroviruses encode dUTPase. J. Virol. **66:**1791–1794.
4. **Gadsen, M. H., E. M. McIntosh, J. C. Game, P. J. Wilson, and R. H. Haynes.** 1993. dUTP pyrophosphatase is an essential enzyme in Saccharomyces cerevisiae. EMBO J. **12:**4425–4431.
5. **Hopp, T. P., and K. R. Woods.** 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. USA **78:**3824–3828.
6. **Kwong, P. D., R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson.** 1998. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing antibody. Nature **393:**648–659.
7. **Larsson, G., L. A. Svensson, and P. O. Nyman.** 1996. Crystal-structure of the Escherichia coli dUTPase in complex with a substrate analog. Nat. Struct. Biol. **3:**532–538.
8. **Lerner, D. L., P. C. Wagaman, T. R. Phillips, O. Prosperogarcia, S. J. Henriksen, H. S. Fox, F. E. Bloom, and J. H. Elder.** 1995. Increased mutation frequency of feline immunodeficiency virus lacking functional deoxyuridine-triphosphatase. Proc. Natl. Acad. Sci. USA **92:**7480–7484.
9. **Lopez, F., C. Abergel, and J.-M. Claverie.** The LookFor program. http://igs-server.cnrs-mrs.fr/lookfor.
10. **Los Alamos HIV Sequence database.** http://hiv-web.lanl.gov/.
11. **McGeoch, D. J.** 1990. Protein sequence comparisons show that the 'pseudoproteases' encoded by poxviruses and certain retroviruses belong to the deoxyuridine triphosphate family. Nucleic Acids Res. **18:**4105–4110.
12. **McIntosh, E. M., D. D. Ager, M. H. Gadsen, and R. H. Haynes.** 1992. Human dUTP pyrophosphatase—cDNA sequence and potential biological importance of the enzyme. Proc. Natl. Acad. Sci. USA **89:**8020–8024.
13. **Prasad, G. S., E. A. Stura, D. E. McRee, G. S. Laco, C. Hasselkuslight, J. H. Elder, and C. D. Stout.** 1996. Crystal structure of dUTP pyrophosphatase from feline immunodeficiency virus. Protein Sci. **5:**2429–2437.
14. **Rizzuto, C. D., R. Wyatt, N. Hernandez-Ramos, Y. Sun, D. Kwong, W. A. Hendrickson, and J. Sodroski.** 1998. A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. Science **280:**1949–1953.
15. **Sharp, P. M., D. L. Robertson, F. Gao, and B. H. Hahn.** 1994. Origins and diversity of human immunodeficiency viruses. AIDS **8:**S27–S42.
16. **Steagall, W. K., M. D. Robek, S. T. Perry, F. J. Fuller, and S. L. Payne.** 1995. Incorporation of uracil into viral-DNA correlates with reduced replication of EIAV in macrophages. Virology **210:**302–313.
17. **Strahler, J. R., X. X. Zhu, N. Hora, Y. K. Wang, P. C. Andrews, N. A. Roseman, J. V. Neel, L. Turka, and S. M. Hanash.** 1993. Maturation stage and proliferation-dependent expression of dUTPase in human T-cell. Proc. Natl. Acad. Sci. USA **90:**4991–4995.
18. **Turelli, P., F. Guigen, J.-F. Mornex, R. Vigne, and G. Quérat.** 1997. dUTPase-minus caprine arthritis-encephalitis virus is attenuated for pathogenesis and accumulates G-to-A substitutions. J. Virol. **71:**4522–4530.