Volume 250, number 2, 591-595 FEB 07344 July 1989

# Patterns of nucleotide substitutions and implications for the immunological diversity of human immunodeficiency virus

Nobuaki Shimizu, Takashi Okamoto*, Etsuko N. Moriyama+, Yasuhiro Takeuchi, Takashi Gojobori+ and Hiroo Hoshino

*Department of Hygiene, Gunma University School of Medicine, Maebashi, Gunma-ken 371, *Virology Division, National Cancer Center Research Institute, Tsukiji, Chuo-ku, Tokyo 104* and *+Department of Evolutionary Genetics, National Institute of Genetics, Mishima, Shizuoka-ken 411, Japan*

Received 30 March 1989

Human immunodeficiency virus (HIV) exhibits immunological hypervariability, which has been an obstacle to successful production of effective anti-HIV vaccines. In this study, we estimated patterns of nucleotide and amino acid substitutions in the *env* gene of HIVs, with the aim of finding characteristics of the mechanism which generates the immunological diversity of the *env* protein of HIVs. We found that nucleotide changes between A and G are predominant compared to those between other nucleotides. Since this feature is consistent with the pattern of nucleotide substitutions of other retroviral genes but is quite different from those of most eukaryotic genes, a high rate of nucleotide substitution between A and G appears to be specific for retroviruses including HIVs. We discuss the biological relationship between this biased substitution and the mechanism generating hypervariability of epitopes on the *env* protein of HIVs.

Human immunodeficiency virus; Immunological diversity; Nucleotide substitution

## 1. INTRODUCTION

Human immunodeficiency virus (HIV) is thought to be the causative agent of acquired immune deficiency syndrome (AIDS) and AIDS-related complex [1-5]. It is well known that HIV shows a great immunological diversity which is an obstacle to the production of effective anti-HIV antibodies in a patient [6,7]. In recent years, nucleotide sequences of HIV genomes of various isolates have been determined [7-9,16]. In comparing such sequences with each other, many nucleotide substitutions have been particularly observed on *env* genes [10]. The high rate of nucleotide substitution gives rise to great immunological diversity in the *env* protein in HIVs. To clarify the molecular mechanism of generating HIV variations, we estimated the pattern of nucleotide and amino acid substitutions for the *env* gene of HIVs isolated in various locations in the world.

Here, we show that for the *env* gene of HIVs, the number of nucleotide substitutions between A and G is much larger than for those of other substitutions. Moreover, the high rate of nucleotide substitution between A and G tends to cause changes in hydrophilic amino acids much more frequently than in hydrophobic amino acids. Since this feature of nucleotide substitutions is quite different from those of non-retroviral genes, it appears to be specific to retroviruses [12]. This specific pattern of nucleotide substitutions may play an important role in generating the immunological hypervariability of the *env* protein in HIVs.

*Correspondence address:* T. Gojobori, Department of Evolutionary Genetics, National Institute of Genetics, Mishima 411, Japan

## 2. MATERIALS AND METHODS

To evaluate a mutation pattern for the *env* gene of HIVs, we used the nucleotide sequences of six molecular clones of HIV genomes, *BRU* [8], *BH8*, *HXB3*, *HXB2*, *BH10* [7] and *p.v.22* [9]. Since these clones were all derived from patients in New York, they are considered to be evolutionarily close to each

*Published by Elsevier Science Publishers B.V. (Biomedical Division)*

other. In fact, the phylogenetic tree of HIVs suggests that they have diverged from a common ancestor within a decade [10]. For this reason, almost all nucleotide changes between two isolates compared can be attributed to single nucleotide substitutions.

We compared nucleotide and amino acid sequences between the following pairs of isolates: *BRU/BH8*, *BH8/HXB3*, HXB3/HXB2, *HXB2/BH10*, *BH10/p.v.22*, and *HXB2/p.v.22*. Because these pairs showed the closest relationships in the phylogenetic tree, these comparisons were made to avoid counting multiple nucleotide substitutions as much as possible. We then counted the number of nucleotide changes between the pairs A/G, A/C, A/T, G/C, G/T, and T/C, separately. The frequency of nucleotide changes is given by $p(X,Y) = n_{XY}/(n_X + n_Y)$, where $n_{XY}$ represents the number of nucleotide changes between nucleotides X and Y, and $n_X$ and $n_Y$ are the total numbers of nucleotides X and Y, respectively (X or Y = A, T, G or C). Subsequently, the relative frequency of nucleotide changes between X and Y is obtained from $f_{XY} = p(X,Y)/\sum_{X \neq Y} p(X,Y)$.

Similarly, we also determined the number of amino acid changes within: hydrophilic, hydrophobic, neutral amino acids; and between: hydrophilic/hydrophobic, hydrophilic/neutral, hydrophobic/neutral amino acids [14,15]. The frequency of amino acid changes is also given by $n_{AB}/(n_A + n_B)$, where $n_{AB}$ is the number of amino acid changes between two hydropathic characteristics A and B (A, B = 'hydrophilic', 'hydrophobic', or 'neutral'), and $n_A$ and $n_B$ are the total numbers of amino acids which have hydropathic characteristics A and B, respectively.

These frequencies were compared separately for the entire *env* gene and its coding regions for the extracellular and transmembrane domains of the *env* protein. Furthermore, nucleotide changes were counted separately at the 1st, 2nd and 3rd positions of codons.

Table 1

Relative frequencies of six types of nucleotide changes at the 1st, 2nd and 3rd positions of codons in the entire *env* gene and in the extracellular and transmembrane domains of the *env* protein of HIVs

| Nucleotide change | Codon position | | | |
|---|---|---|---|---|
| | 1st | 2nd | 3rd | All |
| A ↔ T | 15.0 (5)[a] | 8.1 (5) | 3.3 (3) | 7.0 (13) |
| A ↔ C | 19.1 (6) | 19.5 (10) | 9.2 (7) | 13.5 (23) |
| A ↔ G | 54.7 (21) | 36.0 (20) | 43.9 (37) | 42.8 (78) |
| G ↔ T | 7.3 (2) | 0.0 (0) | 2.6 (2) | 2.0 (4) |
| G ↔ C | 3.9 (1) | 16.1 (7) | 15.5 (10) | 13.9 (18) |
| C ↔ T | 0.0 (0) | 20.2 (10) | 25.5 (18) | 20.7 (28) |

| | Domain of the *env* protein | |
|---|---|---|
| | Extracellular | Transmembrane |
| A ↔ T | 10.5 (11) | 2.5 (2) |
| A ↔ C | 16.1 (16) | 10.4 (7) |
| A ↔ G | 41.3 (41) | 45.5 (37) |
| G ↔ T | 0.0 (0) | 4.2 (4) |
| G ↔ C | 7.1 (5) | 21.9 (13) |
| C ↔ T | 25.0 (18) | 15.4 (10) |

[a] Numbers in parentheses show the absolute numbers of nucleotide changes. Values expressed as %

## 3. RESULTS

Table 1 lists the relative frequencies of nucleotide changes between A/G, A/C, A/T, G/C, G/T, C/T for the entire region of the *env* gene and its coding regions for the extracellular and transmembrane domains of the *env* protein of HIVs. Throughout these regions, we counted the number of nucleotide changes at the 1st, 2nd and 3rd codon positions separately. It should be pointed out that for all three codon positions in the whole *env* region, the rates of nucleotide substitution between A/G were always found to be the highest followed by those of C/T. For the 2nd and 3rd codon positions, the rates of nucleotide substitution between A/G are approx. 2-fold greater than those of T/C. As most of the nucleotide substitutions at the 3rd codon position do not cause alterations in the amino acids, the pattern of nucleotide changes at that position can reflect the pattern of mutation at the nucleotide level.

At the 1st codon position in the *env* gene, the rate of nucleotide changes between A/G is the highest but no nucleotide change between C/T was found. The absence of nucleotide changes between C/T seems to be attributable to random errors by a low frequency of nucleotide substitutions at the 1st codon position, because its relative frequency would become 3.9% if only one nucleotide change between C/T had occurred.

Table 2 shows the proportions of nucleotide changes at the 1st, 2nd and 3rd codon positions in

Table 2

Proportions (%) of nucleotide changes at the 1st, 2nd and 3rd codon positions for the *env* gene of HIVs

| Codon position | Domain of the *env* protein | | Entire *env* region |
|---|---|---|---|
| | Extracellular | Transmembrane | |
| 1st | 26.4 | 15.5 | 21.2 |
| 2nd | 38.0 | 28.2 | 33.8 |
| 3rd | 35.6 | 56.3 | 45.0 |

the *env* gene. For the coding regions of the extracellular domain of the *env* protein, the proportion of nucleotide changes at the 2nd codon position is the highest among all codon positions. This feature suggests that the functional constraint against amino acid changes is imposed very weakly on the extracellular domain of the *env* protein, since all nucleotide substitutions at the 2nd codon position lead to amino acid changes. On the other hand, for the coding regions of the transmembrane domain of the *env* protein, the proportion of nucleotide changes at the 3rd codon position is much higher than those at the other two codon positions. This feature suggests that the functional constraint against amino acid changes is strongly imposed on the transmembrane domain of the *env* protein of HIVs, since most of the nucleotide substitutions at the 3rd codon positions are synonymous (silent). This feature also indicates that nucleotide substitution between A/G can be generated much more frequently than other types of substitutions of HIV genomes.

In table 3, we show the relative frequencies of changes among hydrophilic, hydrophobic and neutral amino acids for the *env* protein of HIVs. For the extracellular domains of the *env* protein, the changes within hydrophilic amino acids and between hydrophilic/neutral amino acids occur at much higher rates compared to other types of amino acid changes. For the transmembrane domains of the *env* protein, the changes within hydrophobic amino acids and between hydrophilic/neutral amino acids also occur at much higher rates than other types of amino acid changes. It is thought that most of these three types of amino acid changes do not impair the biological function and structural conformation of the *env* protein. This is because such amino acid changes do not greatly alter the distribution of electric charges on the *env* protein which is important in maintaining the structural conformation. The rate of changes between hydrophilic/hydrophobic amino acids appears to be lower vs other types of amino acid changes in the extracellular and transmembrane domains. These types of amino acid changes often affect the biological function and structural conformation of the *env* protein, because they produce significant alterations in the electric charges on the *env* protein.

In table 4, we show the frequency of four kinds of nucleotides at the three codon positions when all 61 sense codons of the genetic code are classified into three groups of codons encoding hydrophilic, hydrophobic and neutral amino acids. It should be noted that at the 2nd codon position, the codons for hydrophilic amino acids contain only A or G and not C or T. The codons for hydrophobic amino acids, on the other hand, have only C or T and not A or G at the same position. Thus, the nucleotide substitutions between A/G at the 2nd codon position result in changes in hydrophilic rather than hydrophobic amino acids. Furthermore, amino acid changes caused by the nucleotide

Table 3

Relative frequencies of changes within or among hydrophilic, hydrophobic and neutral amino acids (in %)

| Amino acid change[a] | Domain of the *env* protein | | Entire *env* region |
|---|---|---|---|
| | Extracellular | Transmembrane | |
| Philic ↔ Philic | 22.4 | 12.7 | 19.6 |
| Philic ↔ Neutral | 27.0 | 38.2 | 30.8 |
| Phobic ↔ Phobic | 17.8 | 32.4 | 22.4 |
| Phobic ↔ Neutral | 13.8 | 8.8 | 11.9 |
| Philic ↔ Phobic | 13.2 | 7.8 | 11.2 |
| Neutral ↔ Neutral | 5.7 | 0.0 | 4.2 |

Philic, phobic and neutral: hydrophilic, hydrophobic and neutral amino acids, respectively

Table 4

Frequencies of four kinds of nucleotides at the 1st, 2nd and 3rd codon positions when all 61 sense codons of the genetic code are classified into three groups of codons encoding hydrophilic, hydrophobic and neutral amino acids

| Amino acid | Nucleotide | Codon position | | | |
|---|---|---|---|---|---|
| | | 1st | 2nd | 3rd | All |
| Hydrophilic | A | 6/18 | 12/18 | 5/18 | 23/54 |
| | G | 4/18 | 6/18 | 5/18 | 15/54 |
| | C | 8/18 | 0/18 | 4/18 | 12/54 |
| | T | 0/18 | 0/18 | 4/18 | 4/54 |
| Hydrophobic | A | 4/20 | 0/20 | 5/20 | 9/60 |
| | G | 8/20 | 0/20 | 5/20 | 13/60 |
| | C | 4/20 | 4/20 | 5/20 | 13/60 |
| | T | 4/20 | 16/20 | 5/20 | 25/60 |
| Neutral | A | 6/23 | 4/23 | 5/23 | 15/69 |
| | G | 4/23 | 7/23 | 6/23 | 17/69 |
| | C | 4/23 | 12/23 | 6/23 | 22/69 |
| | T | 9/23 | 0/23 | 6/23 | 15/69 |

substitutions between A/G and between C/T (transitions) may not greatly affect the structural conformation and biological function of proteins. This is because, according to the genetic code, nucleotide substitutions between A/G and between C/T never cause changes between hydrophilic/hydrophobic amino acids.

However, epitopes on the *env* protein of HIVs are believed to consist mainly of hydrophilic amino acids. Therefore, if nucleotide substitutions between A/G occur much more frequently than the other types of substitutions in the HIV genomes, they would give rise to significant alterations in the antigenicities of the *env* protein of HIVs.

## 4. DISCUSSION

In this study, we have examined the patterns of nucleotide and amino acid substitutions for the *env* gene of HIVs in order to clarify the mechanism of generation of the immunological diversity of the *env* protein. We have found that nucleotide substitutions between A/G occur much more frequently than other types. Moreover, we observed that according to the genetic code, codons for hydrophilic amino acids have only A and G at the 2nd position where all nucleotide substitutions are nonsynonymous (amino acid altering), and that the codons for hydrophobic amino acids have only C and T at the same position.

Based on these findings, we speculate that the immunological diversity of HIVs is generated during replication of the HIV genome as a result of mutations between A/G occurring much more frequently than other base changes. This is supported by the pattern of nucleotide substitutions for the *gag* gene. In fact, we also observed that for the 3rd codon position of the *gag* gene the relative frequency (60.2%) of nucleotide changes between A/G was the highest of all the types of nucleotide changes, and that it was much greater than the relative frequency (12.3%) of nucleotide changes between C/T which was the second highest. It was also found that for the frequency of nucleotide substitutions at the 3rd codon position between viral and cellular oncogenes, the change between A/G was much more frequent than other types of changes [12]. For influenza viruses, on the other hand, the rate of nucleotide substitutions between A/G is comparable to that between C/T [13]. The HIV genomes and all viral oncogenes have a reverse transcriptional step in their replication cycles whereas the influenza viruses do not. Recently, low fidelity of the reverse transcription of HIV has been reported [11]. For this reason, we speculate that the reverse transcription of retroviruses including HIVs causes nucleotide changes between A/G much more frequently than other types of changes.

Since the codons for hydrophilic amino acids have only A and G at the 2nd position, the nucleotide substitutions between A/G can result in higher rates of change in hydrophilic compared to hydrophobic amino acids for the *env* protein of HIVs. Since epitopes on the *env* protein of HIVs consist mainly of hydrophilic amino acids, a high rate of nucleotide substitution between A/G can markedly alter the antigenicities of epitopes on the *env* protein without affecting the distribution of electric charges in it. Thus, the frequent nucleotide substitution between A/G may help to generate the immunological hypervariability of the *env* protein in HIVs.

In this report, we have shown the pattern of nucleotide and amino acid substitutions in the *env* gene of HIVs and tried to explain the mechanism for generating the immunological diversity of HIVs. We believe that our findings will provide insights into studies not only on the molecular evolution of HIV genomes, but also on the development of effective anti-HIV drugs and vaccines.

*Acknowledgement:* This study was supported by a Grant-in-Aid from the Ministry of Education, Science and Culture of Japan.

## REFERENCES

[1] Barré-Sinoussi, F., Chermann, J.C., Rey, F., Nugeyre, M.T., Chamaret, S., Gruest, J., Danguet, C., Axler-Blin, C., Vezinet-Brun, F., Rouziox, C., Rozenbaum, W. and Montagnier, L. (1983) Science 220, 868–870.

[2] Gallo, R.C., Salahuddin, S.Z., Popovic, M., Shearer, G.M., Kaplan, M., Haynes, B.F., Palker, T.J., Redfield, R., Oledke, J., Safai, B., White, G., Foster, P. and Markham, P.D. (1984) Science 224, 500–503.

[3] Popovic, M., Sarngadharan, M.G., Read, E. and Gallo, R.C. (1984) Science 224, 497–500.

[4] Sarngadharan, M.G., Popovic, M., Bruch, L., Schubpack, J. and Gallo, R.C. (1984) Science, 223, 506–508.

[5] Schupbach, J., Popovic, M., Gilden, R.V., Gonda, M.A., Sarngadharan, M.G. and Gallo, R.C. (1984) Science 224, 503–508.

[6] Rabson, A. and Martin, M.A. (1985) Cell 40, 477–480.
[7] Ratner, L., Haseltine, W., Patarca, R., Livak, K.J., Starcich, B., Josephs, S.F., Doran, E.R., Rafalski, J.A., Whitwhoen, E.A., Baumeister, K., Ivanoff, L., Patteway, S.R., Pearson, M.L., Lautenberger, J.A., Papas, T.S., Ghrayeb, J., Chang, N.T., Gallo, R.C. and Wong-Staal, F. (1985) Nature 313, 277–284.
[8] Wain-Hobson., Sonigo, P., Danos, O., Cole, S. and Alizon, M. (1985) Cell 40, 9–17.
[9] Muesing, M.A., Smith, D.H., Cabradilla, C.D., Benton, C.V., Lasky, L.A. and Capon, D.J. (1985) Nature 313, 450–458.
[10] Yokoyama, S., Moriyama, N.E. and Gojobori, T. (1987) Proc. Jap. Acad. 63, 147–150.
[11] Takeuchi, Y., Nagumo, T. and Hoshino, H. (1988) J. Virol. 62, 3900–3902.
[12] Gojobori, T. and Yokoyama, S. (1988) J. Mol. Evol. 26: 148–156.
[13] Saitou, N. (1987) Jap. J. Genet. 62, 439–443.
[14] Kyte, J. and Doolittle, R.F. (1982) J. Mol. Biol. 157, 105–132.
[15] Hopp, T.P. and Woods, K.R. (1981) Proc. Natl. Acad. Sci. USA 78, 3824–3828.
[16] Myers, G., Rabson, A.B., Josephs, S.F., Smith, T.F. and Wong-Staal, F. (1988) Human Retroviruses and AIDS, Los Alamos National Laboratory, N.M.