Amino Acids (1998) 14:343–352

Amino
Acids

Analysis of HIV by entropy evolution rate

K. Sato[1], **S. Miyazaki**[2] and **M. Ohya**[1]

[1] Department of Information Sciences, Science University of Tokyo, Noda City, Chiba, Japan
[2] Center for Information Biology, National Institute of Genetics, Mishima City, Shizuoka, Japan

Accepted October 29, 1997

Summary. We analyze the variation of HIV after infection by means of an information measure, called the entropy evolution rate. In our analysis, we use a part of the external glycoprotein gp120 including the V3 region observed from six patients.
Then we could make the following two aspects clear;
(1) the relation between the change of the entropy evolution rate and the appearance of symptoms of disease, and
(2) the relation between the change of the entropy evolution rate and that of the CD4 count of the patients.

Keywords: Amino acids – HIV – Entropy evolution rate

1. Introduction

The main purpose of this study is to find a new criterion grasping the processes of the change of CD4 count and the immunity of patients from the gene level after HIV infection.

In Section 2, we summarize the data of HIV genes of the six patients used in this paper. The entropy evolution rate and the method of how to use it for our analysis are discussed in this section. In Section 3, we present our results by the graphs. We discuss our results and the usefulness of our method in Section 4.

2. Material and methods

We consider two aligned amino acid (resp. base) sequences $\mathcal{A}$ and $\mathcal{B}$, which are composed of 20 (resp. 4) kinds of amino acids (resp. bases) and the gap *. The complete event system $(\mathcal{A}, p)$ of $\mathcal{A}$ is determined by the occurrence probability p_i of each amino acid (resp. base) a_i and the gap * $(0 \leq i \leq 20)$ (resp. $0 \leq i \leq 4$) with $a_0 = *$;

$$\begin{pmatrix}\mathcal{A}\\ p\end{pmatrix}=\begin{pmatrix}*, a_1,\ldots, a_{20}\\ p_0, p_1,\ldots, p_{20}\end{pmatrix}\quad\left(\text{resp.}\begin{pmatrix}*, a_1,\ldots, a_4\\ p_0, p_1,\ldots, p_4\end{pmatrix}\right)$$

In the same way, the complete event system $(\mathcal{B},q)$ of $\mathcal{B}$ is

$$\begin{pmatrix}\mathcal{B}\\ q\end{pmatrix}=\begin{pmatrix}*, a_1,\ldots, a_{20}\\ q_0, q_1,\ldots, q_{20}\end{pmatrix}\quad\left(\text{resp.}\begin{pmatrix}*, a_1,\ldots, a_4\\ q_0, q_1,\ldots, q_4\end{pmatrix}\right)$$

We can construct the compound event system $(\mathcal{A}\times\mathcal{B}, r)$ for two sequences $\mathcal{A}$ and $\mathcal{B}$.

$$\begin{pmatrix}\mathcal{A}\times\mathcal{B}\\ r\end{pmatrix}=\begin{pmatrix}(**), (*a_1),\ldots, (a_{20}a_{20})\\ r_{00}, r_{01},\ldots, r_{2020}\end{pmatrix}\left(\text{resp.}\begin{pmatrix}(**), (*a_1),\ldots, (a_4a_4)\\ r_{00}, r_{01},\ldots, r_{44}\end{pmatrix}\right),$$

where r_{ij} represents the joint probability of the event i of $\mathcal{A}$ and the event j of $\mathcal{B}$.

These event systems define various entropies, among which the following two are important:

(1) Shannon entropy

$$S(\mathcal{A})=-\sum_i p_i\log p_i,$$

which expresses the amount of information carried by $(\mathcal{A},p)$.

(2) The mutual entropy

$$I(\mathcal{A},\mathcal{B})=\sum_{i,j} r_{i,j}\log\frac{r_{i,j}}{p_iq_j},$$

which expresses the amount of information transmitted from $\mathcal{A}$(resp. $\mathcal{B}$) to $\mathcal{B}$(resp. $\mathcal{A}$).

Using the above information measures, a measure indicating the difference between two amino acid sequences was introduced in (Ohya, 1989). This measure is called the entropy evolution rate and defined as follows: Put

$$r(\mathcal{B}/\mathcal{A})=\frac{I(\mathcal{A},\mathcal{B})}{S(\mathcal{A})},$$

which is the rate how much information is transmitted from $\mathcal{A}$ to $\mathcal{B}$, and it is symmetrized as

$$r(\mathcal{A},\mathcal{B})=\frac{1}{2}\{r(\mathcal{A}/\mathcal{B})+r(\mathcal{B}/\mathcal{A})\}$$

The entropy evolution rate $\varrho(\mathcal{A},\mathcal{B})$ is defined by

$$\varrho(\mathcal{A},\mathcal{B})=1-r(\mathcal{A},\mathcal{B})$$

In this paper, we use this entropy evolution rate to examine the variation of HIV sequences of six patients. The entropy evolution rate takes the value in $[0, 1]$; $\varrho(\mathcal{A},\mathcal{B})=0$ if $\mathcal{A}$ and $\mathcal{B}$ are completely same and $\varrho(\mathcal{A},\mathcal{B})=1$ if they are completely different. Therefore the variation of HIV becomes larger, the entropy evolution rate is getting larger.

Data used in our analysis are the base sequences of HIV for the six patients reported in (Wolfs et al., 1991; Holmes et al., 1992; McNearney et al., 1992). We obtained the data from the International Nucleotide Sequence Database (DDBJ/EMBL/GenBank). Here, the six patients are designated as patient A to patient F. The facts reported for the six patients are summarized in Table 1.

Table 1. Data used in our analysis

Designation in our analysis	Patient A	Patient B	Patient C	Patient D	Patient E	Patient F
Designation in our analysis	Patient 1	Patient 495	Patient 82	s 1	s 2	s 4
Presumed trans mission mode	homosexual contact	homosexual contact	a single batch of factor VIII	no information	no information	no information
Clinical status	p24antigenemia (1988)	AIDS (1989) p24antigenemia	asymptomatic	no information	no information	no information
CD4 counts	decreasing	decreasing	decreasing	fluctuating	decreasing	decreasing
Antiviral therapy	None	AZT (1989)	None	None	None	None
Term	1985– (about 5 years period)	1985– (about 5 years period)	1984–1991 (7 years period)	1985.11–89.5 (4.5 years period)	1985.5–87.10 (2.5 years period)	1985.1–89.6 (4.5 years period)
Length	183–276nt	183–276nt	234nt	332–335nt	332–335nt	332–335nt
Tissue	serum	serum	plasma	peripheral blood leucocyte	peripheral blood leucocyte	peripheral blood leucocyte
Molecular type	RNA	RNA	RNA	DNA	DNA	DNA

We use the sequence, a part of the gp120 region including the third variable (V3) region whose mutation rate is particularly high in HIV (Watson et al., 1993; de Jong et al., 1992). The V3 region is composed of disulfide bounds of cysteine residues located in the amino acids 296 and 330 of the gp120, as shown in Fig. 1. Although it has been called the principal neutralization domain (PND) as the antibody for the segment is able to block the HIV infection, the antibody for a specific virus is gradually losing its effect because of the variation (mutation) of the virus.

It is reported (Wolfs et al., 1991) that patient B was diagnosed as having AIDS in about 5 years after the primary infection and that patient A remained healthy although the p24-antigenemia reappeared at 3 years after his infection. The p24-antigenemia in blood reflects the amount of the virus, it is reappeared when the patient has AIDS, so that it is used as a value measuring the condition of the patient.

The CD4 count has been used as a measure to know having AIDS by several researcher and medical doctors. The number of CD4 for patient D fluctuates, and that for other patients gradually decreases. This CD4 count represents the number of immunocyte destroyed by HIV. The immunocyte for healthy people is around from 800 to 1,000 per $1\mu\ell$-blood. When the CD4 count of a patient decreases and it becomes less than 200, various infections are considered to appear. Therefore, according to the diagnosis standard of CDC (Centers for Disease Control), when the CD4 become less than 200, the patient is recognized to have AIDS. The CD4 count is reported only for patient D, E, and F. For patient D, it changes as 470, 826, 273, and 515 from the primary stage (called 1 year) up to the fourth stage (called 4 year). For the patient E and F, it changes as 1225, 756, 368 and 943, 575, 187, respectively.

The number of the sequence data observed from the six patients are listed in Table 2.

We used the base sequences having the same length of bases for each patient. For example, in the primary stage of patient A, we used 6 data out of 8 data because the length of six data is 276 and that of other two is 183. Moreover, in order to carry out our analysis, first we translate the base sequences of HIV collected from the six patients into the amino acid sequences, and secondly, we directly used the base sequences. Our analysis is done in the following two cases (I) and (II).

(I) In order to compare the genome sequences of HIV in successive years, the entropy evolution rate is computed for the sequences obtained at one year with respect to those obtained at the next year (we call it the entropy evolution rate for each year), and we examine the variation (mutation) rate with the mean of the entropy evolution rates for each year and their standard deviation for each year.

(II) In order to check the variation of HIV from the primary stage, we compute the entropy evolution rate for the sequences of each year w.r.t. the primary year (we call it the entropy evolution rate for the primary year). Similarly as the case (I), we examine the variation by means of the entropy evolution rates for the primary year and their standard deviations.

As an example, we explain how to compute the entropy evolution rate and others mentioned above in (I) and (II) for the patient A. From Table 2, the number of genome sequences for the patient A are as follows: n = 6 (year 0), n = 7 (year 1), n = 7 (year 2), n = 5 (year 3), n = 6 (year 4), n = 6 (year 5). For the case (I), we compute every entropy evolution rate for the aligned sequences in successive years, for instance, $\varrho(A_i^3, A_j^4)$ (i = 1,

IVIRSDNITDNAKTIIVQLKEAVQIN CTRPNNNTRKSIHIGPGKAFYATGEIIGDIRQAHC NLSRVDWEDTLKQIAEKLREQFRNKTIVFNQ

IVIRSDNITDNSKTIIVQLKEAVQIN CTRPNNNTRKSIHIGPGKAFYATGEIIGDIRQAHC NLSRVDWEDTLKQIAEKLREQFRNKTIVFNQ

IVVRSDNITDNAKTIIVQLKKAVQIN CIRPNNNTRKSIHIGPGKAFYATGETIGDIRQAHC NLSGGDWENTLKQIAEKLREQFRNKTIVFNQ

Fig. 1. 3 sequence data collected from patient A in primary stage of infection. It's V3 region is underlined

Table 2. The number of sequences used in this paper

Patient A	year 0	year 1	year 2	year 3	year 4	year 5
Number of data collected from GenBank	8	7	9	9	9	8
Number of data used	6	7	7	5	6	6
Patient B	year 0	year 1	year 2	year 3	year 4	year 5
Number of data collected from GenBank	11	6	6	6	7	8
Number of data used	7	3	4	4	4	4
Patient C	year 0	year 3	year 4	year 5	year 6	year 7
Number of data collected from GenBank	1	15	11	23	15	13
Number of data used	1	15	11	23	15	13
Patient D	year 0	year 2	year 3	year 4		
Number of data collected from GenBank	5	2	4	3		
Number of data used	5	2	4	3		
Patient E	year 0	year 2	year 2.5			
Number of data collected from GenBank	5	5	6			
Number of data used	5	5	6			
Patient F	year 0	year 4	year 4.5			
Number of data collected from GenBank	5	6	6			
Number of data used	5	6	6			

$\ldots, 5.\ j = 1, \ldots, 6)$ for the sequence A_i^3 of the third year and the sequence A_j^4 of the fourth year. Then we compute their mean value given by

$$\bar{\varrho}\left(A^3, A^4\right) \equiv \frac{\sum_{i=1}^{5}\sum_{j=1}^{6}\varrho\left(A_i^3, A_j^4\right)}{30},$$

which enables us to examine the variation of HIV. In the same way, we compute $\bar{\varrho}(A^0,A^1)$, $\bar{\varrho}(A^1,A^2)$, $\bar{\varrho}(A^2,A^3)$, $\bar{\varrho}(A^3,A^4)$, $\bar{\varrho}(A^4,A^5)$. The standard deviation of the entropy evolution rate for year 3 and year 4 is defined as follows:

$$\sqrt{\frac{\sum_{i=1}^{5}\sum_{j=1}^{6}\left\{\varrho\left(A_i^3, A_j^4\right) - \bar{\varrho}\left(A^3, A^4\right)\right\}^2}{30}}$$

For the case (II), we compute the mean entropy evolution rates for every sequence of each year with respect to that of the primary year. For instance, the mean entropy evolution rate for the fifth year w.r.t. the primary year is given by

$$\bar{\varrho}\left(A^0, A^5\right) = \frac{\sum_{i=1}^{6}\sum_{j=1}^{6}\varrho\left(A_i^0, A_j^5\right)}{36}.$$

We similarly compute $\bar{\varrho}(A^0,A^1)$, $\bar{\varrho}(A^0,A^2)$, $\bar{\varrho}(A^0,A^3)$, $\bar{\varrho}(A^0,A^4)$, $\bar{\varrho}(A^0,A^5)$, and their standard deviations.

All six patients are examined with these quantities, and our results are shown in the next section.

Here we note that we should align the sequences to compute the entropy evolution rate, and the alignment is done by the method in (Ohya and Uesaka, 1970; Neeleman and Wunsch, 1970).

3. Results

The following figure (Fig. 2) is the results of the mean entropy evolution rates for each year and the standard deviations obtained from the amino acid

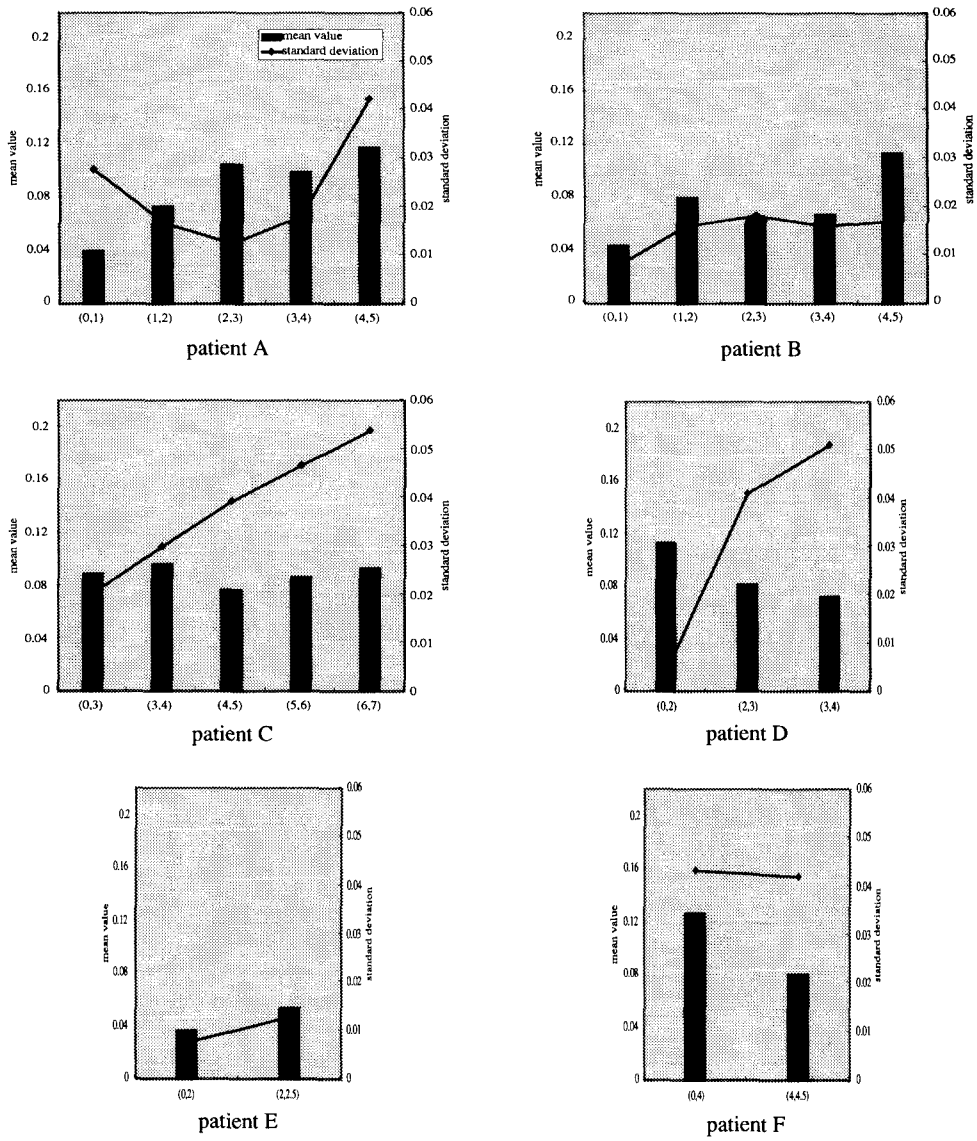

Fig. 2. Mean entropy evolution rate (bars) and standard deviation (lines) for each year

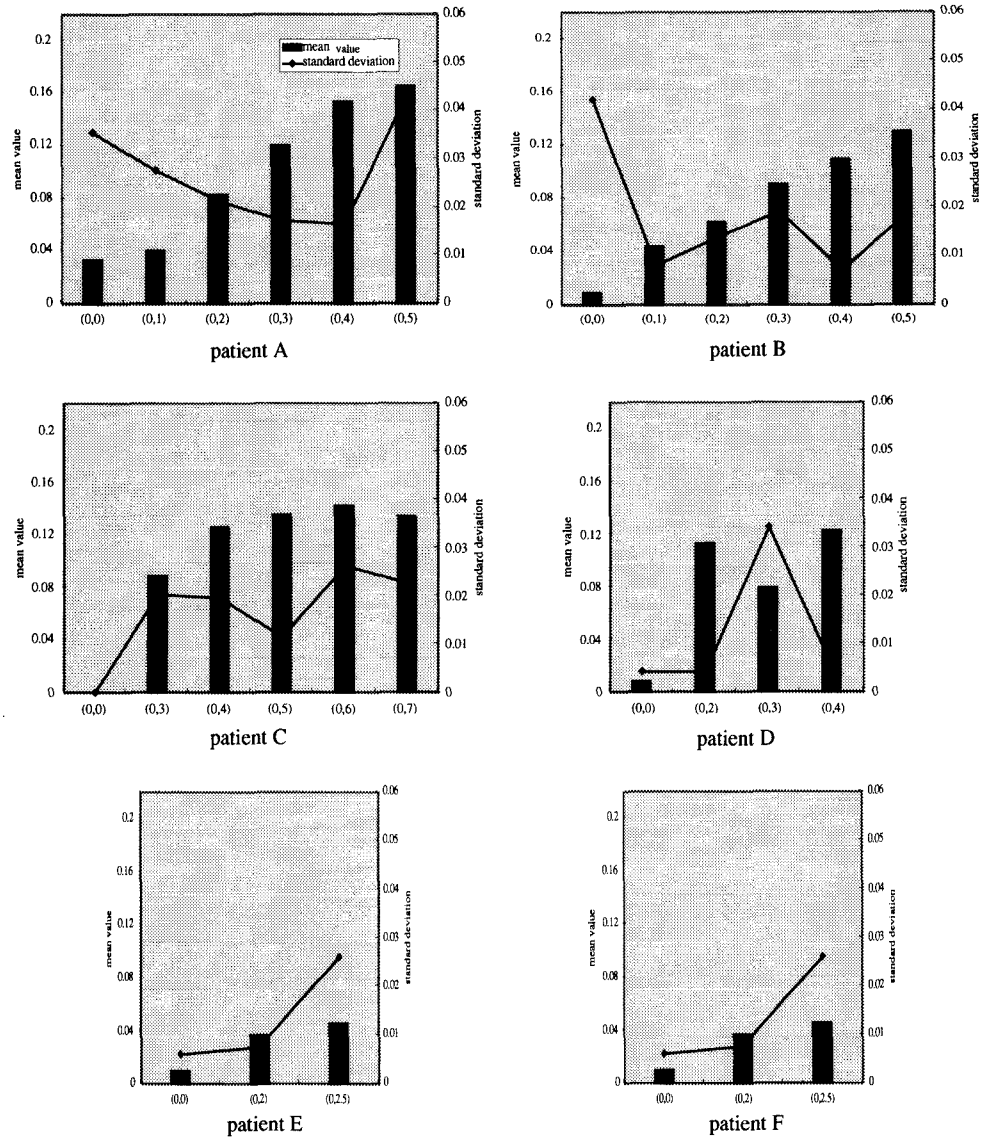

Fig. 3. Mean entropy evolution rate (bars) and standard deviation (lines) measured from primary year

sequences. Here $(i, i + 1)$ denotes the $(i + 1)$-th year w.r.t. i-th year and the mean value means the mean entropy evolution rate.

Fig. 3 shows the results of the mean entropy evolution rates for the primary year and their standard deviations, so that $(0, i)$ denotes the i-th year w.r.t. the primary year.

The following figures (Fig. 4, Fig. 5) are the results obtained from the base sequences.

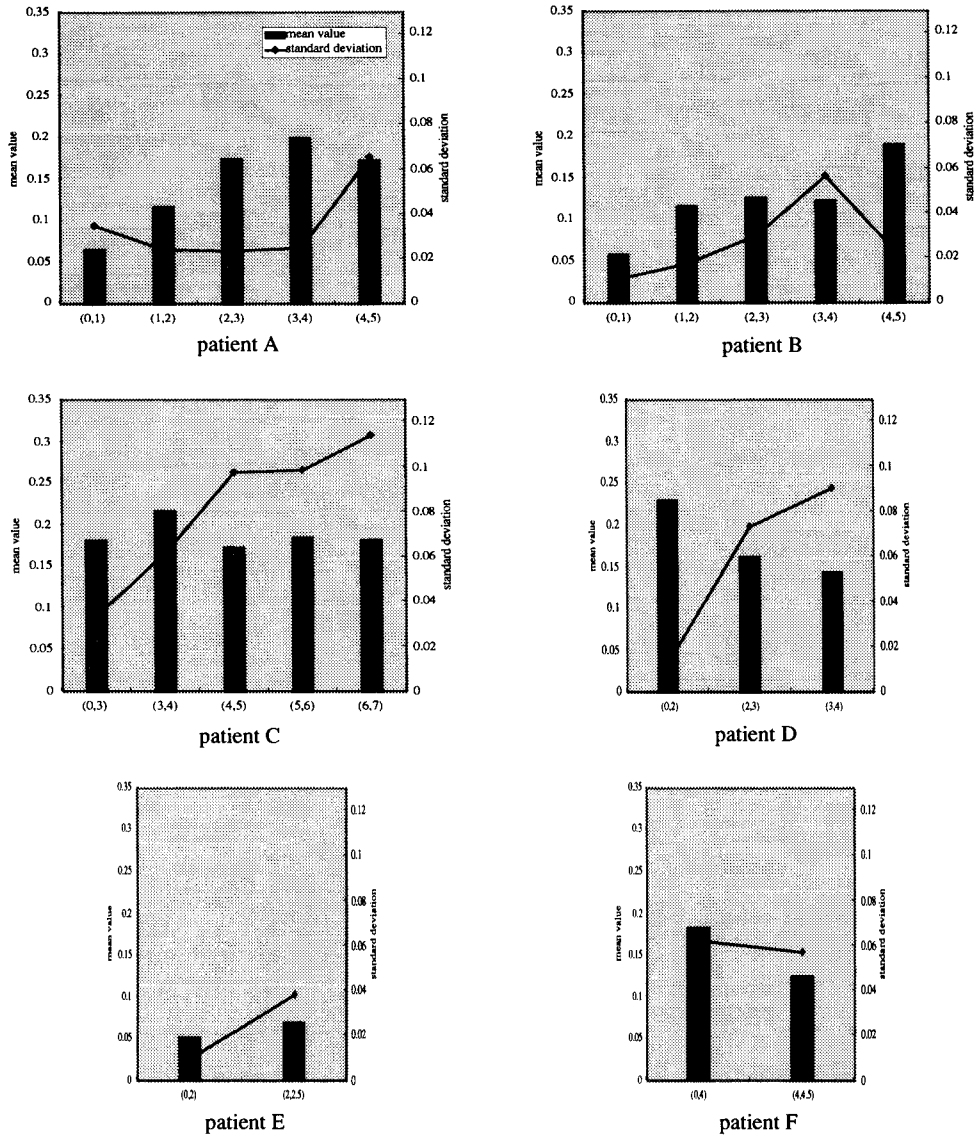

Fig. 4. Mean entropy evolution rate (bars) and standard deviation (lines) for each year

4. Discussion

Patient B was diagnosed as having AIDS at about 5 years after the primary infection. According to the result of the mean entropy evolution rate (m-EER for short) for each year, the change of the m-EER for the patient B is met the second extreme increase at that time. The change of the m-EER (Fig. 2) for patient B is considered as a fundamental pattern of the outbreak of AIDS. Based on this pattern, we may say the following conclusions for other patients. Patient A will be diagnosed as having AIDS in a few years, because the second extreme increase seems starting. Similarly, patient C will have an attack of AIDS soon, because the second moderate increase is occurred.

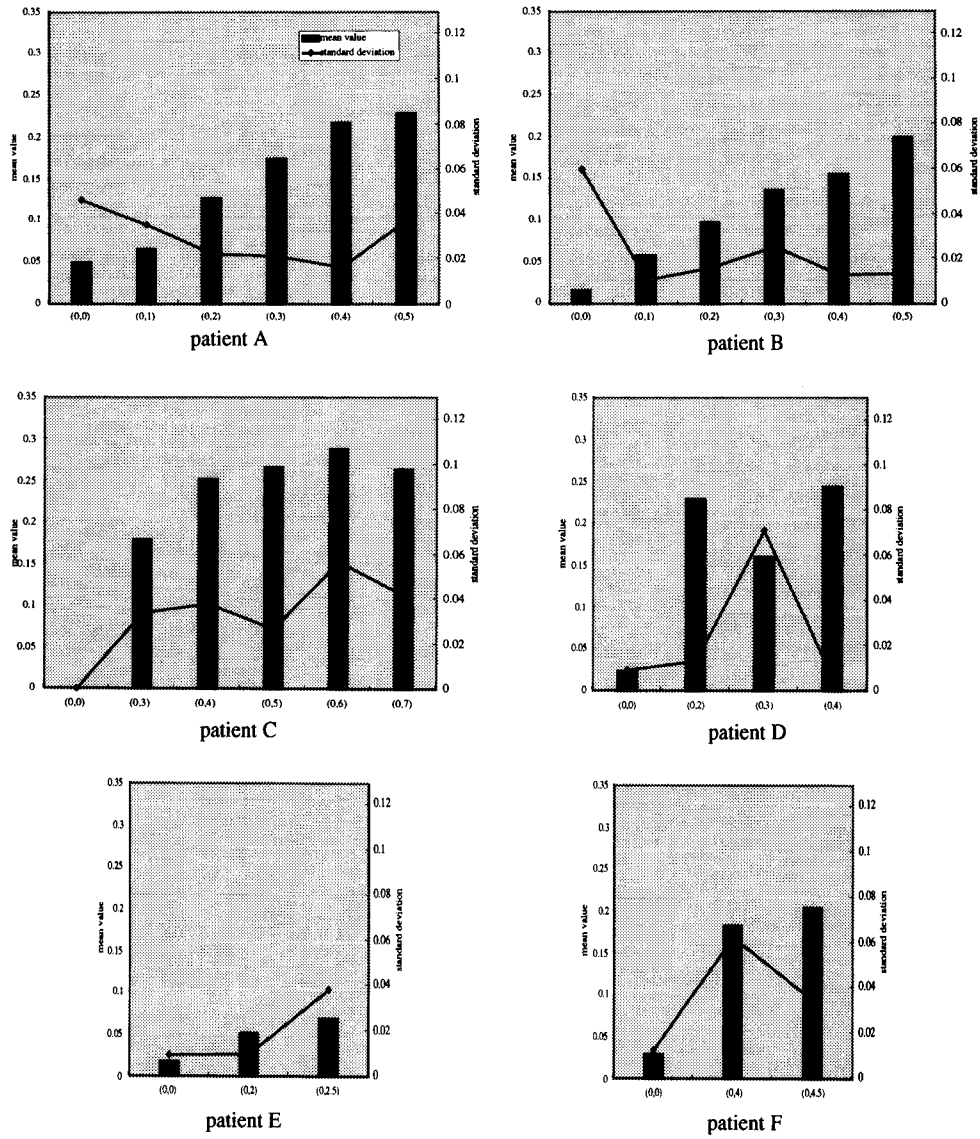

Fig. 5. Mean entropy evolution rate (bars) and standard deviation (lines) measured from primary year

Patient D, patient E and patient F have few number of data, so that we merely say a few comments. Patient D may possibly increase here after. Patient E may be far from the outbreak of AIDS. Since the data of patient F are lacked a first few years, it is very difficult to judge the variation of his HIV.

According to the result of the mean entropy evolution rate measured from the primary stage, the m-EERs of the patients except the patient D increase as shown in Fig. 3. This consequence agrees with a report that the CD4 counts of the patients except patient D decrease. That is, the gradual decrease of

the CD4 count is equal to the increase of the m-EER for the primary year. Further, the CD4 count of patient D fluctuates and his m-EER also fluctuates. This result means that there exists a positive correlation between the m-EER for the primary year and the CD4.

We merely note that the standard deviation shows how many different HIV exist in each stage (year).

The results obtained by using the whole base sequences are shown in Fig. 4 for the case (I) and Fig. 5 for the case (II). These results show that patient B, like the case of the amino acid sequences, have a characteristic variation and other patients are also similar to the case of the amino acid sequences.

From our analysis, we may conclude that the mean entropy evolution rate can be a measure of the variation of HIV and the outbreak of AIDS as the CD4 count.

References

Holmes EC, Zhang LQ, Simmonds P, Ludlam CA, Brown AJL (1992) Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. Evolution 89: 4835–4839

de Jong JJ, Goudsmit J, Keulen W, Klaver B, Krone W, Tersmette M, de Ronde A (1992) Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity. J Virol 66: 757–765

McNearney T, Hornickova Z, Markham R, Birdwell A, Arens M, Saah A, Ratner L (1992) Relationship of human immunodeficiency virus type 1 sequence heterogeneity to stage of disease. Medical Sciences 89: 10247–10251

Needleman SB, Wunsch CD (1970) A general method applicable to search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443–453

Ohya M (1989) Information theoretical treatment of genes. Trans IEICE E 725: 556–560

Ohya M, Uesaka Y (1992) Amino acid sequences and DP matching: new method for alignment. Information Sciences 63: 139–151

Watson JD, Gilman M, Witkowski J, Zoller M (1993) Recombinant DNA, 2nd edn. Freeman and Company

Wolfs TW, Zwart G, Bakker M, Valk M, Kuiken C, Goudsmit J (1991) Naturally occurring mutations within HIV-1 V3 genomic RNA lead to antigenic variation dependent on a single amino acid substitution. Virology 185: 195–205

Authors' address: Dr. Keiko Sato and Prof. Dr. M. Ohya, Department of Information Sciences, Science University of Tokyo, Noda City, Chiba 278, Japan.

Received May 29, 1997