Analysis of HIV by Entropy Evolution Rate

Keiko SATO* • Satoru MIYAZAKI*1 • Masanori OHYA*2

Abstract

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 variation of the entropy evolution rate and the appearance of symptoms of disease, and
(2) the relation between the variation of the entropy evolution rate and that of the CD4 count of the patients.

Key Words : Amino Acid, 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 six patients used in this paper. The entropy evolution rate and the method 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 A and B, which are composed of 20 (resp. 4) kinds of amino acids (resp. bases) and the gap*.

The complete event system A of 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 $a_0 = *$;

$$\{A\} = \\left\{ a_0, a_1, \ldots, a_\alpha \right\} (\text{resp. } \left\{ p_0, p_1, \ldots, p_\alpha \right\})$$

In the same way, the complete event system (B, q) of B is

$$\{B\} = \\left\{ a_0, a_1, \ldots, a_\beta \right\} (\text{resp. } \left\{ q_0, q_1, \ldots, q_\beta \right\})$$

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

$$\{A \times B\} = \\left\{ a_0, a_1, \ldots, a_\alpha a_0, a_1, \ldots, a_\alpha \right\} (\text{resp. } \left\{ r_0, r_1, \ldots, r_\alpha \right\})$$

where $r_i$ represents the joint probability of the event $i$ of A and the event $j$ of B.

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

(1) Shannon entropy

$$S(A) = - \sum p_i \log p_i,$$

which expresses the amount of information carried by $(A, p)$. 

内容は「Amino Acids (1998) 14:343-352」に発表されたものを少し変更して使用。

* Department of Control and Computer Engineering
*1 Center for Information Biology, National Institute of Genetics, Mishima City, Shizuoka, Japan
*2 Department of Information Sciences, Science University of Tokyo, Noda City, Chiba, Japan
(2) The mutual entropy

\[ I(A, B) = \sum_i \log \frac{p(i,j)}{p(i)} \]

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

Using the above information measures, a measure indicating the difference between two amino acid sequences was introduced in (1). This measure is called the entropy evolution rate and defined as follows: Put

\[ r(B/A) = \frac{I(A, B)}{S(A)} \]

which is the rate how much information is transmitted from \( A \) to \( B \), and it is symmetrized as

\[ r(A, B) = \frac{1}{2} (r(A/B) + r(B/A)) \]

The entropy evolution rate \( p(A, B) \) is defined by

\[ p(A, B) = 1 - r(A, B) \]

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 \([0,1]\); \( p(A, B) \) if \( A \) and \( B \) are completely the same and \( p(A, B) = 1 \) 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 (2-4). 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

<table>
<thead>
<tr>
<th>Sequence type</th>
<th>Patient A</th>
<th>Patient B</th>
<th>Patient C</th>
<th>Patient D</th>
<th>Patient E</th>
<th>Patient F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transmission route</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>Transmission from</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>CDR evolution</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>CDR evolution location</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>CD4 count</td>
<td>( &lt; 200 )</td>
<td>( &lt; 200 )</td>
<td>( &lt; 200 )</td>
<td>( &lt; 200 )</td>
<td>( &lt; 200 )</td>
<td>( &lt; 200 )</td>
</tr>
<tr>
<td>Viruses</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>Marker</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>CD4 count of patient</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>CD4 count of patient D</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>CD4 count of patient E</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>CD4 count of patient F</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>DNA</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>RNA</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

We use the sequence, a part of the gp120 region including the third variable (V3) region whose mutation rate is particularly high in HIV (5, 6). 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 the effect because of the variation (mutation) of the virus.

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

It is reported (2) 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 decrease. This CD4 count represents the number of immunocyte destroyed by HIV. The immunocyte for healthy people is around from 800 to 1000 per 1 μl -blood. When the CD4 count of a patient decreases and it becomes less than 200, the various infections are considered to appear. Therefore, according to the diagnosis standard of CDC (Centers for Disease Control), when the CD4 becomes 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
Analysis of HIV by Entropy Evolution Rate

515 from the primary stage (1 year) up to the fourth stage (4 year). For 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.

Table 2. The number of sequences used in this paper

<table>
<thead>
<tr>
<th>Patient</th>
<th>year 0</th>
<th>year 1</th>
<th>year 2</th>
<th>year 3</th>
<th>year 4</th>
<th>year 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient A</td>
<td>3 6 9 8 9 9 8</td>
<td>3 6 8 8 4 8 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient B</td>
<td>1 6 6 6 7 8</td>
<td>1 6 6 6 7 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient C</td>
<td>1 15 11 15 13 13</td>
<td>1 3 9 15 13 13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient D</td>
<td>1 2 4 3</td>
<td>1 2 4 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient E</td>
<td>5 6 6</td>
<td>5 6 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient F</td>
<td>5 6 6</td>
<td>5 6 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 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 the 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 (mutation) rate with the mean 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 patient A. From Table 2, the number of genome sequences for 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, $p(A_i, A_i')$ $(i = 1, \cdots, 5; j = 1, \cdots, 6)$ for the sequence $A_i'$ of the third year and the sequence $A_i$ of the fourth year. Then we compute their mean value given by

$$\bar{p} = \frac{\sum_{i=1}^{5} \sum_{j=1}^{6} p(A_i', A_i)}{30}$$

which enables us to examine the variation of HIV. In the same way, we compute $\bar{p}(A_i, A_i')$, $\bar{p}(A_i, A_i')$, $\bar{p}(A_i, A_i')$, $\bar{p}(A_i', A_i')$, $\bar{p}(A_i', A_i')$. The standard deviation of the entropy evolution rate for year 3 and year 4 is defined as follows:

$$\sqrt{\frac{\sum_{i=1}^{5} \sum_{j=1}^{6} (p(A_i', A_i))^2}{30}}$$

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

$$\bar{p}(A_i', A_i') = \frac{\sum_{i=1}^{5} \sum_{j=1}^{6} p(A_i', A_i')}{36}$$

We similarly compute $\bar{p}(A_i', A_i')$, $\bar{p}(A_i', A_i')$, $\bar{p}(A_i', A_i')$, $\bar{p}(A_i', A_i')$, 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 (7, 8).

3. Results

The following figures (Fig.2) is the results of the mean entropy evolution rates for each year and the standard deviations obtained from the amino acid sequences. Here \((i, j)\) denotes the \((j+1)\)-th year w.r.t. \(i\)-th year and the mean value means the mean entropy evolution rate.

![Fig.2](image_url)

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

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

Since the results obtained from the base sequences was the almost same as the amino acids sequences, we omit them in this research report.

4. Discussion

Patient B was diagnosed as having AIDS 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 variation of the m-EER for patient B has met the second extreme increase at this time. The variation 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 can give the following conclusions for other patients. Patient A will be diagnosed as having AIDS in a few years, because the
Analysis of HIV by Entropy Evolution Rate

second extreme increase seems to be starting. Similarly, patient C will have an attack of AIDS soon, because the second moderate increase has occurred. Patient D, patient E and patient F have only a few number of data, so we can merely make 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 has lacked in the 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-EER of the patients except patient D increases as shown in Fig.3. This consequence agrees with a report that the CD4 count of the patients except patient D decreases. 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 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.

Reference

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