数理解析研究所講究録
1013巻 1997年 92-101

エントロピー進化率によるHIV変化過程の解析

Masanori Ohya　　Keiko Sato
東京理科大学情報科学科　　大矢雅則　・　佐藤圭子

1. はじめに

ウイルスの変化，とくにHIVの変化を情報理論をベースとして導入されたエントロピー進化率と呼ばれる尺度を用いて解析する．こうした解析を通して，この尺度がHIVに感染した患者のCD4値，免疫等の変化の過程を遺伝子レベルから把握する指標の一つになることがわかる．

本論文の構成は次のようになっている．2節では，この論文の解析的基礎となる遺伝子のエントロピー進化率について説明する．3節では，解析する患者のHIV遺伝子のデータについて述べ，4節ではそのデータを第2節で説明した解析方法に適用する．5節では，解析結果をグラフ化して示し，6節で得られた結果を次の観点から考察する．

(1) エントロピー進化率の変化と患者のAIDS発症との関わり．
(2) エントロピー進化率の情報的尺度の変化と患者のCD4値の変化との関連．

2. エントロピー進化率

アライメントによって得られた2つのアミノ酸配列 $\mathcal{A},\mathcal{B}$ について，構成要素である20種のアミノ酸の出現確率をそれぞれ p_i,q_j $(1\leq i,j\leq 20)$ とし，ギャップ*の出現確率を p_0,q_0 とすると，完全事象系は次のようになる．

$$\begin{pmatrix}\mathcal{A}\\ p\end{pmatrix}=\begin{pmatrix}*, & A, & \mathrm{L}, & Y\\ p_0, & p_1, & \mathrm{L}, & p_{20}\end{pmatrix}$$

$$\begin{pmatrix}\mathcal{B}\\ q\end{pmatrix}=\begin{pmatrix}*, & A, & \mathrm{L}, & Y\\ q_0, & q_1, & \mathrm{L}, & q_{20}\end{pmatrix}$$

さらに配列 $\mathcal{A}$ の各アミノ酸と配列 $\mathcal{B}$ の各アミノ酸が同時に生起する確率を $r_{i,j}$ とすると，複合完全事象系は，

$$\begin{pmatrix} \mathcal{A}\times\mathcal{B} \\ r \end{pmatrix} = \begin{pmatrix} **, & *A, & \text{L}, & YY \\ r_{00}, & r_{01}, & \text{L}, & r_{2020} \end{pmatrix}$$

となる.

そこで，これらの完全事象系のもとに遺伝子配列における各種エントロピーが以下のように定められる.

(1) Shannonエントロピー

$$S(\mathcal{A}) = -\sum_{i=0}^{20} p_i \log p_i$$

これは，配列$\mathcal{A}$のもつ情報量を表している.

(2) 相互エントロピー

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

この相互エントロピーは，$\mathcal{A}$と$\mathcal{B}$との間での情報のやりとりの精度を表すもので，$\mathcal{A}$（または$\mathcal{B}$）から$\mathcal{B}$（または$\mathcal{A}$）へ伝えられた情報量と考えることができる.

上記の2つの量をもとに，以下の量が導入されている.

(3) エントロピー比

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

これは，$\mathcal{A}$がもつ情報量と$\mathcal{A}$から$\mathcal{B}$へ伝達された情報量の比であり，$\mathcal{A}$に対する$\mathcal{B}$の類似度を表す尺度である.

(4) 対称エントロピー比

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

$\mathcal{A}$と$\mathcal{B}$の両方からみたエントロピー比の平均をとったものを対称エントロピー比といい，$\mathcal{A}$と$\mathcal{B}$の類似度を示している.

この対称エントロピー比をベースとして遺伝子配列$\mathcal{A}$と$\mathcal{B}$との相違を示す量$\rho(\mathcal{A},\mathcal{B})$が，

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

で定められる．これがエントロピー進化率［1］である．今回の解析には，患者から取り出したHIVの遺伝子配列の年月による変異の値を調べるため，このエントロピー進化率を用いた．この値は $0 \leq \rho(\mathcal{A},\mathcal{B}) \leq 1$ であり，変異率が大きいとその値は大きくなるものである．

3. HIVデータ

この解析で用いたデータは，［2,3,4］で報告されている6人の患者から採取されたHIVの遺伝子配列である．私たちの解析では，この6人の患者を，それぞれ患者A,B,C,D,E,Fと呼ぶ．

配列データはいずれも，HIVの中でも特に変異率の高いenv領域の外皮糖蛋白gp120の一部分を用いている[5,6]．HIV感染は免疫応答としてgp120を中和する抗体を作り出すが，ウイルスの変異のために特定のウイルスに効く抗体が役に立たなくなってしまう．このgp120の抗原性のある領域は，gp120の296番目と330番目の2つのシステイン(c)に囲まれたV3領域である．中和抗体はこのV3領域に結合する．この解析で用いたすべてのデータにはV3領域が含まれている(Fig.1).

IVIRSDNITDNAKTIIVQLKEAVQIN CTRPNNNTRKSIHIGPGKAFYATGEIIGDIRQAHC NLSRVDWEDTLKQIAEKLREQFRNKTIVFNQ

IVIRSDNITDNSKTIIVQLKEAVQIN CTRPNNNTRKSIHIGPGKAFYATGEIIGDIRQAHC NLSRVDWEDTLKQIAEKLREQFRNKTIVFNQ

IVVRSDNITDNAKTIIVQLKKAVQIN CIRPNNNTRKSIHIGPGKAFYATGETIGDIRQAHC NLSGGDWENTLKQIAEKLREQFRNKTIVFNQ

Fig.1. 3 sequence data collected from patient A in early (0) stage of infection.
V3 region is the underlined part.

報告されている6人の患者の状態については，Table 1 に要約した．表からもわかるように，患者Bは初期感染から5年目にAIDSと診断されている．また，患者Aは血清中のp24抗原が感染後3年目に再び現れている．血中のp24抗原量はウイルス量を反映し，AIDSを発症する頃から再び検出されるので，HIV感染者の病態を知る一手段として用いられる．

医師などや多くの研究者たちは，AIDS発症を知る手段としてCD4値を目安に使っている．そのCD4値は、患者Dだけが変動し，他の5人の患者は徐々に減少している．このCD4値はHIVが感染する免疫細胞の個数を表す．免疫細胞の個数は普通の人では，血液マイクロリットル中800から1300程度である．感染者ではこれが徐々に減り200を切ると様々な感染症にかかりやすくなり，AIDSに発病するといわれている．CDC（米国・国立防疫センタ）の診断基準によると，200に満たない場合には，かりに

AIDSの症状がなくてもAIDS患者と認定されることになっている．CD4リンパ球数が正確に示されているのは，患者D,E,Fだけである．患者Dは，初期には470，2年目には826，3年目には273，4年目には515とCD4値が変動している．患者Eの採取した年のCD4値は，徐々に1225, 756, 368と減少しており，患者Fも，943, 575と減少し，感染から4年半経って187と低い値になっている．

Table 1. Data used for our analysis

Designation in our analysis	patient A	patientB	patientC	patientD	patientE	patientF
Designation in the original paper	patient1	patient495	patient82	s1	s2	s4
Presumed trans mission mode	homosexual contact	homosexual contact	a singl batch of factor Ⅷ	no information	no information	no information
Clinical status	p24 antigenemia (1988)	AIDS (1989)	asymptomatic	no information	no information	no information
CD4 counts during the study	decreasing	decreasing	decreasing	fluctuating	decreasing	decreasing
Antiviral therapy	None	AZT (1989)	None	None	None	None
Term for the study	1985～ (abaut 5years period)	1985～ (abaut 5years period)	1984～1991 (7years period)	1985.11～89.5 (4.5years period)	1985.5～87.10 (2.5years period)	1985.1～89.6 (4.5years period)
Length	183～276nt	183～276nt	234nt	332～335nt	332～335nt	332～335nt
Tissue	serum	serum	plasma	peripheral blood lelucocyte	peripheral blood lelucocyte	peripheral blood lelucocyte
Molecular type	RNA	RNA	RNA	DNA	DNA	DNA

4. 解析方法

6人の患者から，それぞれ初期感染時から何年間かにわたり採取した配列のデータ数をTable 2 にまとめると，次のようになる．

Table 2. The number of sequences used in this paper

患者A	0年目	1年目	2年目	3年目	4年目	5年目
GenBankから得たデータ数	8	7	9	9	9	8
解析で使ったデータ数	6	7	7	5	6	7

患者B	0年目	1年目	2年目	3年目	4年目	5年目
GenBankから得たデータ数	11	6	6	6	7	8
解析で使ったデータ数	7	3	4	4	4	4

患者C	0年目	3年目	4年目	5年目	6年目	7年目
GenBankから得たデータ数	1	15	11	23	15	13
解析で使ったデータ数	1	15	11	23	15	13

患者D	0年目	2年目	3年目	4年目
GenBankから得たデータ数	5	2	4	3
解析で使ったデータ数	5	2	4	3

患者E	0年目	2年目	2.5年目
GenBankから得たデータ数	5	5	6
解析で使ったデータ数	5	5	6

患者F	0年目	4年目	4.5年目
GenBankから得たデータ数	5	6	6
解析で使ったデータ数	5	6	6

私たちは，患者各々に対して，塩基配列の長さが極端に異なるものは除いて考えている．例えば患者Aの0年目では，8個のデータ中，長さが極端に違う塩基の個数が183のものは除いて276のものを使っているので6種になる．なお，患者から取り出したHIVの塩基配列をアミノ酸配列に翻訳してから，次の2つの場合(I)，(II)について解析を行う．

(I)HIVの遺伝子配列が前回の年に対してどの程度変化が現れたかを比較するために，採取した年のデータと前回採取した年のデータとのエントロピー進化率を計算した．これを各年毎のエントロピー進化率と呼ぶ）．そして，この各年毎のエントロピー進化率の平均と標準偏差の変異率を調べる．

(II)初期感染の頃に比べてどの程度変化したかを調べるために，0年目を基準にしてエントロピー進化率を計算した．これを初期感染に対するエントロピー進化率と呼ぶ．(I)と同様に，平均と標準偏差の変異率を調べる．

具体的に患者Aにおいて説明をしてみる．Table 2からもわかるように，患者Aは，初期に6つの異なる配列データが得られ，1年後には7つ，2年後には7つ，3年後には5

つ，4年後には6つ，5年後には6つが採取された．

解析(I)では，アライメント［7,8］を行った配列に対して，すべての組み合わせのエントロピー進化率を計算する．例えば，感染から2年後の配列A_i^2と3年後の配列A_j^3のエントロピー進化率$\rho(A_i^2, A_j^3)$ $(i=1,\cdots,7.\ j=1,\cdots,5)$は，35通り求められる．したがって，それらの平均値は次のように与えられる．

$$\overline{\rho}(A^2,A^3) \equiv \frac{\sum_{i=1}^{7}\sum_{j=1}^{5}\rho(A_i^2,A_j^3)}{35}$$

同様な方法で，$\overline{\rho}(A^0,A^1),\overline{\rho}(A^1,A^2),\overline{\rho}(A^2,A^3),\overline{\rho}(A^3,A^4),\overline{\rho}(A^4,A^5)$を計算する．これらは，HIVの変化を調べることができる．また2年目と3年目のエントロピー進化率の標準偏差は，次のように，定義される．

$$\sqrt{\frac{\sum_{i=1}^{7}\sum_{j=1}^{5}\left\{\rho(A_i^2,A_j^3)-\overline{\rho}(A^2,A^3)\right\}^2}{35}}$$

解析(II)では，0年目を基準にしたエントロピー進化率の平均を考える．例えば，感染初期A_i^0と4年後の配列A_j^4のエントロピー進化率の平均は，

$$\overline{\rho}(A^0,A^4) \equiv \frac{\sum_{i=1}^{6}\sum_{j=1}^{6}\rho(A_i^0,A_j^4)}{36}$$

となる．同様に$\overline{\rho}(A^0,A^1),\overline{\rho}(A^0,A^2),\overline{\rho}(A^0,A^3),\overline{\rho}(A^0,A^4),\overline{\rho}(A^0,A^5)$を計算する．

前回の採取時との比較，それから初期感染時とその後の各採取時との比較をこの解析方法を用いて、6人の患者A,B,C,D,E,Fすべてに行っている．この結果を次の節でグラフにまとめた．

5. 結果

次のグラフは，解析(I)(Fig.2)と解析(II)(Fig.3)の結果である．

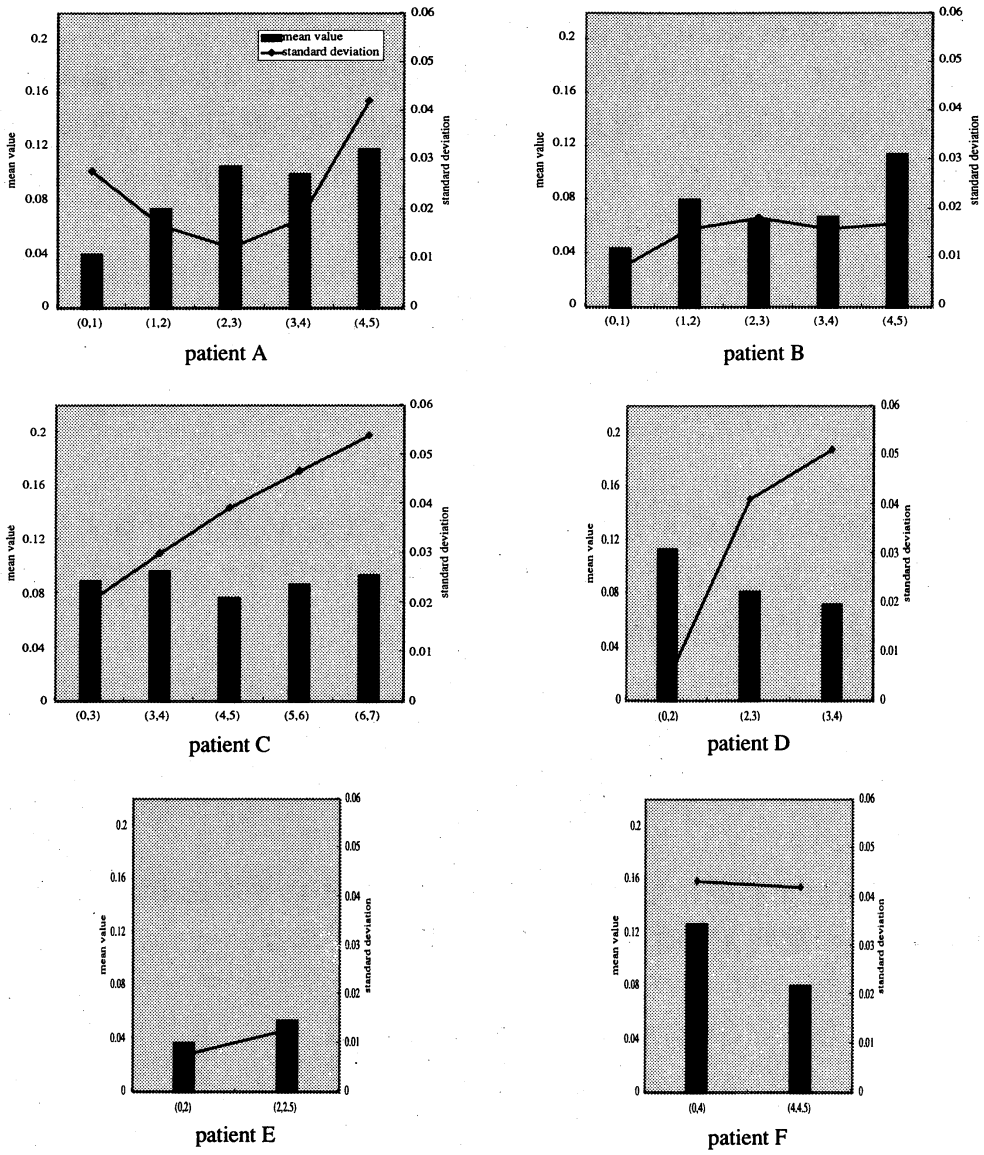

Fig.2. entropy evolution rate (bars) and standard deviation (lines) for each year.

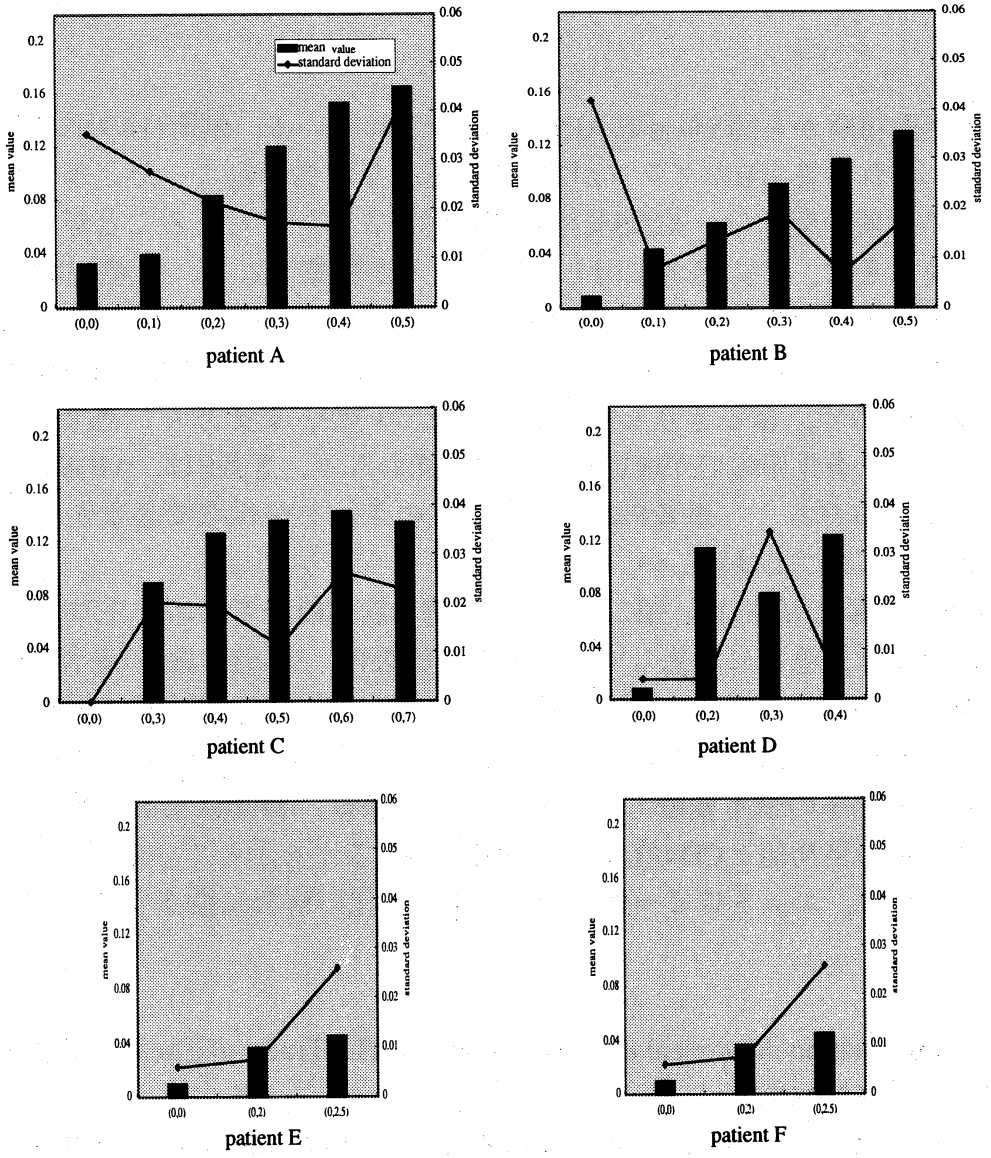

Fig.3. entropy evolution rate (bars) and standard deviation (lines) measured from primary year.

6. 考察

患者Bは感染初期から約5年目にAIDSと診断されている．解析(I)の平均値の結果(Fig.2)から，それは患者Bのグラフが2度目の極端な増加が起こったときであることがわかる．この値の変化は，AIDS発症の基本的パターンと考えることができる．このパターンをもとに，他の患者に対して次のことが推察される．患者Aは2度目のエントロピー進化率の値の増加が起こり始めているので，2,3年後にAIDSと診断されるだろう．同様に，患者Cも患者Aと同じことが言える．また，患者D,E,Fはデータ数が非常に少ないので，はっきりしたことが言えないが，おそらく，患者Dは，その後，エントロピー進化率の値が増加するものと思われる．患者Eは，AIDS発症には，まだ時間がかかるだろう．患者Fは最初の2,3年のデータが採取できなかったので，HIVの変化を調べることは難しい．

解析(II)の平均値の結果(Fig.3)によると，患者D以外の患者のエントロピー進化率の平均値の変化は，徐々に増加している．この結果は，患者D以外の患者のCD4値が徐々に減少しているという報告と反比例する．すなわち，CD4値が徐々減少すると，エントロピー進化率の値が徐々に増加する．さらに，患者DのCD4値が変動すると，エントロピー進化率の値も変動している．この結果は，初期感染を基準にしたエントロピー進化率の平均値とCD4値は正の相関があることを示している．

以上のことから，エントロピー進化率がHIVの変化の解析に役立つことがわかる．

参考文献

[1] M.Ohya, Information theoretical treatment of genes, The Trans. of The IEICE, Vol. E 72, No.5, pp.556-560 (1989)

[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)

[3] E.C. Holmes, L.Q.Zhang, P.Simmonds, C.A. Ludlam, and A.J. L.Brown, Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient, Evolution, Vol. 89, pp.4835-4839 (1992)

[4] T.McNearney, Z.Hornickova, R.Markham, A.Birdwell, M.Arens, A.Saah, and L.Ratner, Relationship of human immunodeficiency virus type 1 sequence heterogeneity to stage of disease, Medical Sciences, Vol. 89, pp.10247-10251 (1992)

[5] J.D.Watson, M.Gilman, J.Witkowski, and M.Zoller, Recombinant DNA Second Edition, Scientific American Books, W.H.Freeman and Company N.Y. (1993)

[6] J.J.de Jong, J.Goudsmit, W.Keulen, B.Klaver, W.Krone, M.Tersmette, and A.de Ronde, Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity, Journal of Virology, pp.757-765 (1992)

[7] M.Ohya and Y.Uesaka, Amino acid sequences and DP matching: A new method for alignment, Information Sciences ¥QTR{bf}{63}, pp.139-151 (1992)

[8] S.B.Needleman and C.D.Wunsch, A general method applicable to search for similarities in the amino acid sequence of two proteins, J.Mol.Biol., pp.443-453 (1970)