2008－BIO－14 (6)
2008／9／18

HIV-1 env 遺伝子 V3 領域の符号構造

矢内 国治　　佐藤 圭子
東京理科大学理工学部情報科学科

概要

HIV-1 (Human Immunodeficiency Virus Type 1) 感染症患者から長期的に採取された env 遺伝子 V3 領域の塩基配列と情報通信における様々な人工的な符号列との近さを測り，その塩基配列が感染症過程をとおしてどのような符号構造を有するのかを調べた．その結果，env 遺伝子 V3 領域の塩基配列は HIV-1 感染症の進行段階に応じて，特定の生成多項式から成る人工符号に近いことがわかった．

The code structure of HIV-1 V3 region

Kuniharu Yanai and Keiko Sato
Tokyo University of Science, Department of Information Sciences

Abstract

We measure the similarity between the nucleotide sequences of V3 region of HIV-1 env gene and those encoded by various artificial codes in information transmission, and we investigate what kinds of code structure the sequences of V3 region have in the course of HIV-1 infection. From this study, we find that in each stage of HIV-1 infection the code structure of the corresponding V3 region is close to each artificial code with a generator polynomial.

1 導入

DNA は 4 つの塩基，アデニン(A)，チミン(T)，シトシン(C)，グアニン(G)から構成されており，1 つの符号列とみなせる．そして DNA の自己複製の際，複製ミス(誤り)がごくまれにしか起きない事実を考えると，DNA や遺伝子は誤りを訂正する能力を備えたそれら特有の符号とみなすことができる．そこで，私たちは，DNA や遺伝子の符号構造を見出すために，まず現在用いられている人工的な誤り訂正符号と DNA の塩基配列の類似を調べ，その塩基配列が人工的な符号によってどこまで説明できるかを調べている[1]．

本研究では遺伝学的差異の 1 つであるエントロピー進化率を用いて，HIV-1 において特に変異しやすい env 遺伝子 V3 領域の塩基配列と人工的な符号を比較し，その塩基配列が感染症過程をとおしてどのような符号構造をもつのか解析を行った．さらに V3 領域における符号構造の変化と患者の病期進行の相関性を調べた．

2 V3 領域の符号論的解析

2．1 遺伝子への符号の適用

遺伝子の符号論的特徴として，次の 3 つが挙げられる．

特徴 1：DNA は 4 つの塩基から構成されている．

特徴 2：塩基の 3 つの組(コドン)が 1 つのアミノ

酸に対応している．

特徴3：遺伝暗号表から，コドンの第3塩基はアミノ酸決定にあまり関与していない．

この3つの特徴から，コドンにおける第1，2塩基は誤り訂正符号の情報記号，第3塩基はその検査記号であると仮説を立てた．これを基に，実際に遺伝子の塩基配列を符号化する際の符号の条件を以下のように定める[2]．

特徴1 → 4元の符号

特徴2 → コドン単位を崩さない符号長

特徴3 → コドンの第3塩基のみ変化する符号

解析では，これらの条件を満たすガロア体 $GF(4)$ 上の $(3k,2k)$符号，$k \in N$ を使用する．それらの符号の特徴を**表1，表2**にまとめる．

なお，塩基 A，T，C，G とガロア体 $GF(4)$ の元は以下のように対応させた．

$$\mathrm{A} \to 0,\ \mathrm{T} \to 1,\ \mathrm{C} \to \alpha,\ \mathrm{G} \to \alpha^2$$

表 1　使用した巡回符号一覧

符号名	略称	符号長	誤り訂正能力	生成多項式の数
(3,2)巡回符号	J(3,2)	3	なし[a]	3
(15,10)巡回符号	J(15,10)	15	なし[a]	51
(21,14)巡回符号	J(21,14)	21	なし[a]	45
(15,10)BCH符号	BCH(15,10)1	15	1 (ランダム誤り)	6
(105,70)BCH符号	BCH(105,70)5	105	5 (ランダム誤り)	12

a. 誤り検出能力のみ

表 2　使用した畳込み符号一覧

符号名	略称	誤り訂正能力	拘束長	符号名	略称	誤り訂正能力	拘束長
自己直交符号1	S1	1 (ランダム誤り)	9	岩垂符号7	I7	21 (バースト誤り)	117
自己直交符号2	S2	1 (ランダム誤り)	27	岩垂符号8	I8	24 (バースト誤り)	132
自己直交符号3	S3	2 (ランダム誤り)	42	岩垂符号9	I9	27 (バースト誤り)	147
自己直交符号4	S4	2 (ランダム誤り)	69	岩垂符号10	I10	3 (バースト誤り)	24
自己直交符号5	S5	3 (ランダム誤り)	120	岩垂符号11	I11	6 (バースト誤り)	39
自己直交符号6	S6	3 (ランダム誤り)	165	岩垂符号12	I12	9 (バースト誤り)	54
岩垂符号1	I1	3 (バースト誤り)	27	岩垂符号13	I13	12 (バースト誤り)	69
岩垂符号2	I2	6 (バースト誤り)	42	岩垂符号14	I14	15 (バースト誤り)	84
岩垂符号3	I3	9 (バースト誤り)	57	岩垂符号15	I15	18 (バースト誤り)	99
岩垂符号4	I4	12 (バースト誤り)	72	岩垂符号16	I16	21 (バースト誤り)	114
岩垂符号5	I5	15 (バースト誤り)	87	岩垂符号17	I17	24 (バースト誤り)	129
岩垂符号6	I6	18 (バースト誤り)	102	岩垂符号18	I18	27 (バースト誤り)	144

2．2　エントロピー進化率

本研究では，様々な符号を用いてHIV-1env遺伝子V3領域を符号化した後，どの符号に近い構造をもつかを調べるための指標として，以下に説明する遺伝学的差異の1つとなるエントロピー進化率を用いる[3]．

今、アライメントによって得られた2つのアミノ酸配列 A と B について考える．完全事象系 (A,p) は 20 種類の各アミノ酸とギャップの生起確率 p_i によって決定される．同様に，完全事象系 (B,q) も q_j によって決定される．

$$(A,p)=\begin{pmatrix} * & A & \cdots & Y \\ p_0 & p_1 & \cdots & p_{20} \end{pmatrix} \qquad p=(p_i)_{i=0}^{20}$$

$$(B,q)=\begin{pmatrix} * & A & \cdots & Y \\ q_0 & q_1 & \cdots & q_{20} \end{pmatrix} \qquad q=(q_j)_{j=0}^{20}$$

さらに，配列 A での各事象と B での各事象に対する同時確率分布 r を求めることにより，複合完全事象系 $(A\times B,r)$ が次のように設定される．

$$(A\times B,r)=\begin{pmatrix} ** & *A & \cdots & YY \\ r_{0,0} & r_{0,1} & \cdots & r_{20,20} \end{pmatrix} \quad r=(r_{i,j})_{i,j=0}^{20}$$

これらの完全事象系から，配列の情報量を示すエントロピーSや，配列間の情報のやりとりの精度を表す相互エントロピーIを求めることができる．

$$S(A)=-\sum_i p_i \log p_i$$

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

相互エントロピーはAとBとの間の情報のやりとりの精度を表すものであるため，この相互エントロピーを用いて類縁度を計ることができる．ここで，AとBとの間の差異を表す量，エントロピー進化率(Entropy evolution rate ; EER) $\rho(A,B)$を，AとBとの情報量を介した相関を表す

$$r(A,B)=\frac{I(A,B)}{S(A)+S(B)-I(A,B)}$$

を用いて，

$$\rho(A,B)=1-r(A,B) \quad (0\le\rho(A,B)\le 1)$$

と定める．$\rho(A,B)$の値が0に近いほど類縁度が高いことになる．

2．3　患者のデータ

解析に用いた患者は16名で，そのデータはHIV sequence databaseより，HIV-1において特に変化の激しいgp120のV3領域の塩基配列を採取した．いずれの患者もHIV-1に感染後，ある一定の間隔でサンプルが採取され，追跡調査が行われた患者である．なお，患者の特徴は**表3**に要約した．

調査期間中，グループ1(A-I)はAIDS発症またはCD4陽性T細胞数200以下に至った患者，グループ2(J-P)はCD4陽性T細胞数200以上を維持し，AIDS発症と診断されなかった患者である．

表 3 患者の要約

患者	抗体陽性転換	AIDS診断(抗体陽性転換後の経過)	死亡	分子タイプ
グループ1				
A	12/1985	6/1990(54ヶ月)	AIDS診断後 39ヶ月	PBMCのウイルスDNA
B	1983-84	(約 6 年)	抗体陽性後 109ヶ月(9.1年)	血漿のウイルスRNAとPBMCのウイルスRNA
C	1983-84	(約 8 年)	抗体陽性後 109ヶ月(9.1年)	血漿のウイルスRNAとPBMCのウイルスRNA
D	1983-84	(約 6.5 年)	抗体陽性後 97ヶ月(8.1年)	血漿のウイルスRNAとPBMCのウイルスRNA
E	1983-84	(約 5 年)	抗体陽性後 85ヶ月(7.1年)	血漿のウイルスRNAとPBMCのウイルスRNA
F	1985	1989(55ヶ月)		血清のウイルスRNA
G	10/1987	11/1990(37ヶ月)		PBMCのウイルスDNA
H	1984	1991(7 年)[a]		血漿のウイルスRNA
I	1983-84	(約 8 年)[a]		血漿のウイルスRNAとPBMCのウイルスRNA
グループ2				
J	1985			血清のウイルスRNA
K	1983-84			血漿のウイルスRNAとPBMCのウイルスRNA
L	1983-84			血漿のウイルスRNAとPBMCのウイルスRNA
M	1984	調査期間中AIDS発症と診断されなかった患者		PBMCのウイルスDNA
N	1985			PBMCのウイルスDNA
O	NA			PBMCのウイルスDNA
P	NA			PBMCのウイルスDNA

a. CD4陽性T細胞数200以下
NAは"Not Available"，PBMCは"末梢血単核細胞"を意味する．

(患者A/G[9]，患者B/C/D/E/I/K/L[4]，患者F/J[5]，患者H[6]，患者M/N[8]，患者O/P[7])

2．4　解析方法

前節で説明したエントロピー進化率を用いて，遺伝子の有する符号と人工的な符号との近さを測る指標を以下のように定める．

①HIV-1感染後のある時期に採取されたV3領域の塩基配列がm本ある場合，それぞれの塩基配列$a_i(i=1,2,\cdots,m)$を符号化する（符号Cによって符号化したものをa_i^cとする）．

②塩基配列a_iと符号化後のa_i^cをそれぞれアミノ酸配列A_iとA_i^cに変換する.

③エントロピー進化率$\rho(A_i, A_i^c)$を計算する. このとき, HIV-1感染後の時期ごとのV3領域の有する符号と人工的な符号Cの平均的な近さを決める指標D_cを以下のように定める.

$$D_c = \frac{\sum_{i=1}^{m} \rho(A_i, A_i^c)}{m}$$

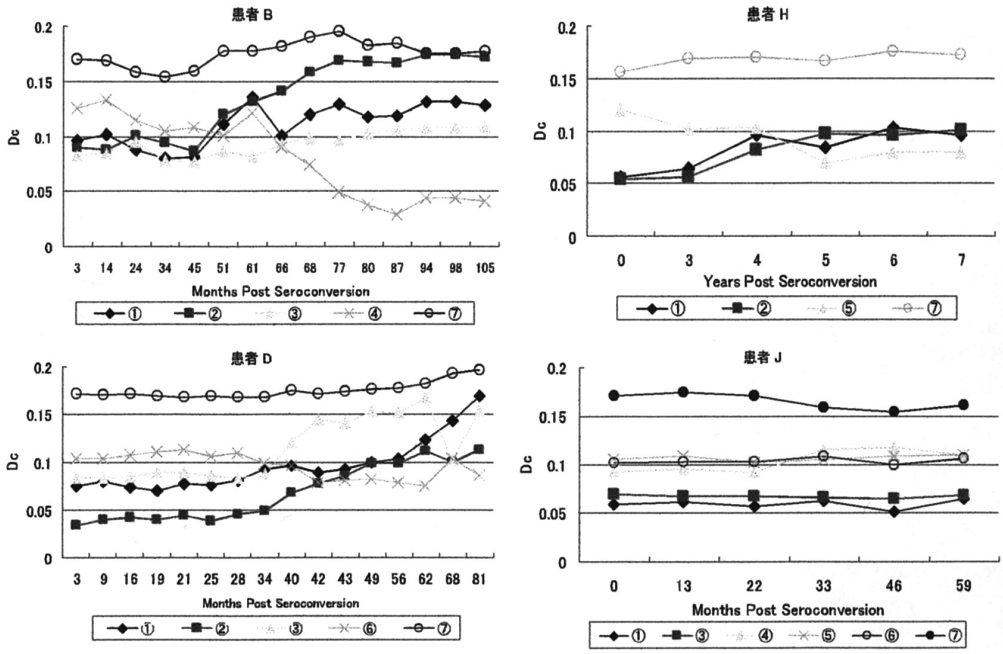
3 結果と考察

HIV-1感染症初期におけるV3領域の塩基配列は, ほとんど全ての患者でJ(15,10)生成多項式$t^5+\alpha^2t^4+t^2+\alpha t+\alpha^2$(グラフ①), J(21,14)生成多項式$t^7+\alpha^2t^6+\alpha t^5+t^4+\alpha^2t^3+\alpha^2$ (②), 自己直交符号S4 (③)の少なくとも2つに近い. しかし, 感染症の進行と共に指標D_cが上昇し, V3領域の符号構造に変化が生じることがわかった.

AIDS発症者の多くで, 感染症が進行するに連れて指標D_cが減少し, V3領域の符号構造に近づく符号がみられた. それは, J(15,10)生成多項式$t^5+t^4+t^2+1$(患者B ④), J(21,14)生成多項式$t^7+\alpha t^5+\alpha t^3+\alpha t^2+\alpha t+1$(患者H ⑤), 自己直交符号S5 (患者D ⑥)のいずれかである.

患者Jのような AIDS未発症者は, 感染初期から継続してJ(15,10)生成多項式$t^5+\alpha^2t^4+t^2+\alpha t+\alpha^2$(①), J(21,14)生成多項式$t^7+\alpha^2t^6+\alpha t^5+t^4+\alpha^2t^3+\alpha^2$(②), 自己直交符号S4 (③)の少なくとも2つに近い. また, J(15,10)生成多項式$t^5+t^4+t^2+1$(④), J(21,14)生成多項式$t^7+\alpha t^5+\alpha t^3+\alpha t^2+\alpha t+1$(⑤), 自己直交符号S5 (⑥)で, AIDS発症者のような指標D_cの減少は見られなかった. その結果を**図1**に示す.

これらの結果より, V3領域の塩基配列はそれ特有の符号構造を有することがわかった. さらにV3領域の塩基配列における符号構造は, HIV-1感染症の病期の進行と共に明らかに変化していくことがわかった. そして, その各進行段階において, V3領域の有する符号に近い符号の生成多項式を特定することができた. J(15,10)とJ(21,14)は誤り訂正能力を持たない. 自己直交符号S4は訂正能力2, S5は訂正能力3である. V3領域がこのような誤り訂正能力がないもしくは低い符号構造に近いことは, この領域の高い変異率と関係があると考えられる.



グラフ番号	符号名	生成多項式	グラフの特徴
①	$J(15,10)$	$t^5+\alpha^2t^4+t^2+\alpha t+\alpha^2$	HIV-1感染症の初期段階で D_c が低い値をとる
②	$J(21,14)$	$t^7+\alpha^2t^6+\alpha t^5+t^4+\alpha^2t^3+\alpha^2$	
③	$S4$	$D^{22}+D^{21}+D^9+D^2+1,$ $D^{38}+D^{35}+D^{10}+D^4+1$	
④	$J(15,10)$	$t^5+t^4+t^2+1$	病期進行と共に D_c が減少
⑤	$J(21,14)$	$t^7+\alpha t^5+\alpha t^3+\alpha t^2+\alpha t+1$	
⑥	$S5$	$D^{39}+D^{36}+D^{23}+D^{21}+D^{14}+1,$ $D^{38}+D^{33}+D^{28}+D^{27}+D^8+1$	
⑦	$J(15,10)$	t^5+1	D_c が高い値をとる (HIV-1感染症過程を通して，この符号構造とは離れた構造をしている)

図 1　各患者の指標 D_c における結果

患者すべてのグラフを表示することは不可能なため，代表としてグループ 1 の患者 B，D，H とグループ 2 の患者 J における V3 領域の各 D_c の値を載せる．

参考文献

[1] K. Sato, N. Fusimi and M. Ohya ; Open Sys. &Information Dyn. 14, 295-306 (2007).

[2] 大矢雅則，松永慎一；電子情報通信学会研究報告，J74-A No. 7, 1075-1084 (1991).

[3] R. S. Ingarden, A. Kossakowski and M. Ohya ; The Transactions Of TheIeice, Vol. E72, 556-560 (1989).

[4] Raj Shankarappa, et al. ; Journal of Virology, Vol. 73, 10489-10502 (1999).

[5] T. W. Wolfs, et al. ; Virology 185, 195-205 (1991).

[6] Edward C. Holmes, et al. ; Proc. Natl. Acad. Sci. USA, Vol. 89, 4835-4839 (1992).

[7] Richard B. Markham, et al. ; Proc. Natl. Acad. Sci. USA, Vol. 95, 12568-12573 (1998).

[8] Steven M. Wolinsky, et al. ; SCIENCE, Vol. 272, 537-542 (1996).

[9] A. B. Wout, et al. ; Journal of Virology, Vol. 72, No. 6, 5099-5107 (1998).