2007－BIO－8 (3)
2007／3／5

インフルエンザA型ウイルスにおけるHA蛋白質のエントロピー型カオス尺度による分類

日谷 尭，佐藤 圭子，大矢 雅則
東京理科大学理工学部情報科学科

概要

本研究では，情報力学におけるカオス尺度の概念を用いて，インフルエンザA型ウイルスにおけるHA蛋白質の進化系統樹作成，分類，変異解析を行った．これにより，従来の分類とは異なる類縁関係の分類やHA蛋白質の進化（変異）の様相を数値的に捉えて特徴付けられることがわかった．

Classification in Influenza A type Virus HA Protein by Entropic Chaos Degree

Takashi HITANI, Keiko SATO and Masanori OHYA
Tokyo University of Science, Department of Information Sciences

Abstract

By means of the entropic chaos degree (ECD) in Information Dynamics, we make evolutionary tree of various Influenza A type virus HA proteins and we analyse and classify evolution of the HA proteins. Then, it is shown that we can obtain new classification of the HA proteins and numerical characterization for their continuous mutation.

1 序論

免疫不全ウイルス，インフルエンザウイルスなどのウイルスは毎年流行を引き起こし，人々の生活を脅かしている．最近では，強病原性のトリインフルエンザウイルスがヒトに感染するタイプに変異するのではないかと世界中で危惧されており，その変異過程の解明が急務となっている．このような中，私達は，ウイルス変異の過程や力学を，情報力学の中で定められたカオスの度合いを計る量的尺度であるエントロピー型カオス尺度（以下，ECD）などの数理とコンピュータ技術を用いて解明することを試みている [2]．既にECDを用いたHIV-1のenv遺伝子の変異解析は行われており，有意義な結果が示されている [9]．本研究では，インフルエンザA型ウイルスにおける赤血球凝集素（以下，HA）蛋白質を用いた解析を行った．

インフルエンザA型ウイルスは複数の蛋白質を持っており，その中で重要な蛋白質の1つが，HA蛋白質である．HA蛋白質はウイルス粒子の表面にある大きな突起状の糖蛋白質であり，抗体に対する抗原となっている．HA蛋白質は，ウイルス粒子が細胞に取り込まれる際に役割を果たし，抗原性の違いによって16種類の亜型（H1～H16）に分類されている．一方，抗原は連続変異と不連続変異という変異で変化することがわかっている．連続変異とは同一の亜型内部での微少な変異のこと，不連続変異とはある亜型が別の種類の亜型に替わることをいう [4][5][6]．

本論文では，ECDとECDとエントロピーの比である$RECD$の生命科学への応用方法について述べた後，ECDを用いた遺伝的差異行列とNJ法によるHA蛋白質の進化系統樹，$RECD$によるHA蛋白質の分類結果，$RECD$による各HA蛋白質の連続変異の解析結果を示す．

2　情報力学の生命科学への応用

ある系の状態 φ（確率分布 p）が力学 Λ によって状態 $\Lambda\varphi$（確率分布 q）に変化するとき，力学 Λ の生成するカオスの度合い $CD(\varphi;\Lambda)$ はカオス尺度と呼ばれている [2][8]．このカオス尺度をエントロピーで表現したものをエントロピー型カオス尺度（以下，ECD）という [7]．本節では，$ECD(p;\Lambda)$ または $ECD(p;\Lambda)$ とエントロピーの比である $RECD(p;\Lambda)$ を用いたアミノ酸配列の変異解析の手法について説明する．まず，配列長が等しい 2 つのアミノ酸配列 X,Y に対して，完全事象系と複合完全事象系を定義する．アミノ酸配列 X の完全事象系 (X,p) は，各アミノ酸 $a_n(n=1,2,\cdots,20)$ の生起確率 p_n とアライメントによって生じたギャップ a_0 の生起確率 p_0 の確率分布 $p=\{p_i\}_{i=0}^{20}$ によって定められる．次に，同様にして，Y の完全事象系 (Y,q) を，アミノ酸とギャップの生起確率による確率分布 $q=\{q_j\}_{j=0}^{20}$ によって，完全事象系 (X,p) と完全事象系 (Y,q) による複合完全事象系 $(X\times Y,r)$ を，a_i と b_j の同時確率分布 $r=\{r_{ij}\}_{i,j=0}^{20}$ によって定める．

$$\begin{pmatrix} X \\ p \end{pmatrix} = \begin{bmatrix} a_0 & , & a_1 & , & a_2 & , & \cdots & , & a_{20} \\ p_0 & , & p_1 & , & p_2 & , & \cdots & , & p_{20} \end{bmatrix}, \begin{pmatrix} Y \\ q \end{pmatrix} = \begin{bmatrix} b_0 & , & b_1 & , & b_2 & , & \cdots & , & b_{20} \\ q_0 & , & q_1 & , & q_2 & , & \cdots & , & q_{20} \end{bmatrix}$$

$$\begin{pmatrix} X\times Y \\ r \end{pmatrix} = \begin{bmatrix} (a_0,b_0) & , & (a_1,b_1) & , & \cdots & , & (a_{20},b_{20}) \\ r_{00} & , & r_{11} & , & \cdots & , & r_{2020} \end{bmatrix}$$

最後に，配列 X と配列 Y における $ECD(p;\Lambda)$ または $RECD(p;\Lambda)$ を次式で求める．（以下，$ECD(p;\Lambda)$ を $ECD(X,Y)$，$RECD(p;\Lambda)$ を $RECD(X,Y)$ と表記する．）

$$ECD(X,Y)=\sum_{i,j=0}^{20} r_{ij}\log\frac{p_i}{r_{ij}} \qquad \begin{array}{ll} ECD(X,Y)>0 & \Leftrightarrow Chaos\ State \\ ECD(X,Y)=0 & \Leftrightarrow No\ Chaos\ State \end{array}$$

$$RECD(X,Y)=\frac{ECD(X,Y)}{S(Y)}=\frac{\sum_{i,j=0}^{20} r_{ij}\log\frac{p_i}{r_{ij}}}{-\sum_{j=0}^{20} q_j\log q_j} \qquad \begin{array}{ll} 0<RECD(X,Y)\leq 1 & \Leftrightarrow Chaos\ State \\ RECD(X,Y)=0 & \Leftrightarrow No\ Chaos\ State \end{array}$$

配列 X から配列 Y への変化の力学は，確率分布 p から確率分布 q に対する力学 Λ と呼ばれる写像（遷移確率）によって与えられると考えられるが，多様な変異をするインフルエンザ A 型ウイルスなどの生命において，確かな写像を知ることは難しい．しかしながら，$ECD(X,Y)$ または $RECD(X,Y)$ は，正確な写像を知ることなく配列 X,Y 間の複雑さを計測出来る．

3　分類と変異解析

本研究で用いたインフルエンザ A 型ウイルス HA 蛋白質のアミノ酸配列は，NCBI の Influenza Virus Resource (http://www.ncbi.nlm.nih.gov/genomes/ FLU/Database/multiple.cgi) から収集した．収集条件は，Influenzavirus A，All Hosts，All Countries/Regions，4 (HA)，All Subtypes，Full-length sequences only とし，アミノ酸数：552〜576，3358 本のアミノ酸配列を収集し使用した．HA 蛋白質の内訳は，H1：456 本（1918 年〜2005 年），H2：72 本（1957 年〜2005 年），H3：1500 本（1963 年〜2006 年），H4：69 本（1956 年〜2004 年），H5：661 本（1959 年〜2006 年），H6：166 本（1963 年〜2005 年），H7：197 本（1927 年〜2006 年），H8：7 本（1968 年〜1992 年），H9：153 本（1966 年〜2004 年，全部で 155 本登録されていたが採取時期不明の 2 本は使用せず），H10：21 本（1949 年〜2004 年），H11：17 本（1956 年〜2004 年），H12：9 本（1976 年〜2005 年），H13：13 本（1977 年〜2004 年），H14：1 本（1982 年），H15：7 本（1979 年〜1983 年），H16：7 本（1975 年〜1999 年）である．各配列には，インフルエンザウイルスの命名法によって，型（本研究では A）/分離された動物種/分離場所/分離された年〔A 型ウイルスの亜型（H，N）〕という名前が付けられている．得られた配列に対し，EMBL（European Molecular Biology Laboratory）の Tompson，Higgins，

Gibson らによって開発されたプログラムである ClustalW を用いてマルチプルアライメントを行い，それらを本研究では使用した [10].

3.1 ECD による進化系統樹

距離法を用いた進化系統樹作成のアルゴリズムの 1 つに，NJ 法（Neighbor-Joining method）がある [3]. これらのアルゴリズムでは，進化系統樹を作成する際，遺伝子配列間の遺伝的差異（または遺伝距離）を求めて遺伝的差異行列（または遺伝距離行列）を作成し，それを用いて進化系統樹を作成する．遺伝的差異の計算には置換率などが用いられるが，ECD を用いて遺伝的差異を以下のように定義すると，ECD による進化系統樹を作成出来る．

$$D(ECD(X,Y)) = \frac{1}{2}\left(ECD(X,Y) + ECD(Y,X)\right)$$

このとき，X, Y をそれぞれ Hi,Hj(i,j=1,...,16) とすると

$$D(ECD(\mathrm{Hi},\mathrm{Hj})) = \frac{1}{2}\left(ECD(\mathrm{Hi},\mathrm{Hj}) + ECD(\mathrm{Hj},\mathrm{Hi})\right) \qquad (\mathrm{i},\mathrm{j} = 1,\ldots,16)$$

となる．尚，Hi,Hj(i,j=1,...,16) は HA 蛋白質の亜型を表している．

分離された動物種を Avian，Equine，Human，Swine として，上記の配列からランダムに選択し，それらを用いて ECD と NJ 法による進化系統樹を作成した (Fig.1)（AccessionNo.ABI85231，AAA43345，ABF60577，AAA43221，AAD13572，ABI84663，ABI84981，ABI85240，AAD49000，ABI84626，ABI84556，ABI84446，BAA14338，ABI84453，AAA96134，ABI85221，AAD25303，ABF60580，AAG17429，AAV30836，AAP47821，AAA43164，CAA44429，ABI20826，AAA43185，BAF37221，AAD21156，ABI85000，CAB95857）．同様にして，分離された動物種が異なる H5 の配列をランダムに選択し，それらを用いて系統樹を作成した (Fig.2)．H5 のアミノ酸配列は，登録されている配列の中では最も多い 8 種類の動物種から分離されており，同一の HA 亜型における動物種による差異を考察するのに適していると考え使用した（AccessionNo.CAA30680，ABF58847，ABH09489，ABI36052，AAC32099，AAT70218，AAT72505，AAT70210）．

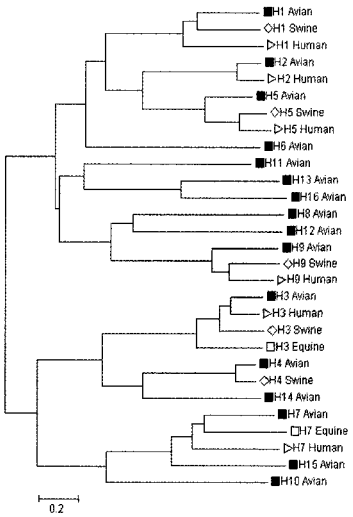

Fig.1 4 種類の動物種から分離された HA の進化系統樹

Fig.1 は，Avian（トリ科），Swine（ブタ），Equine（ウマ科），Human（ヒト）の HA 蛋白質のアミノ酸配列を用いて作成した NJ 法による進化系統樹である．分離された動物種別に，次のような記号を付けて系統樹に表記している（■Avian，◇Swine，□Equine，▷Human）．まず，Fig.1 の最上部の（■H1Avian，◇H1Swine，▷H1Human），（■H2Avian，▷H2Human），（■H5Avian，◇H5Swine，▷H5Human）を見ると，Avian→Swine→Human の順に枝の長さが長くなっていることが見て取れる．これは，Avian に感染するウイルスが，Swine を介して，Human に感染するウイルスに変異したことを表していると考えられる．これは，生物学でわかっている知見と一致する．インフルエンザ A 型ウイルスは，シアル酸をレセプターとして感染するが，このシアル酸は，Avian のものと Human のものでは異なる．しかし，Swine は，Avian のレセプターとなるシアル酸と Human のレセプターとなるシアル酸の両方を保有するいう特異性を持つ．そのため，Swine 内部で，Avian と Human の遺伝子集合体が出来やすい [4]．このことから，この結果は，Avian→Swine→Human の順に HA 蛋白質が変異していったことを示していると言える．

一方，Fig.1 の中央下部分の（■H3Avian，▷H3Human，◇H3Swine，□H3Equine）を見ると，Human までの枝の長さが最も短いことがわかる．この結果は，先に述べた Avian→Swine→Human の順に変異する過程と異なる．しかし，先に述べた Swine の特異性は絶対的なものではないとされている．Avian のウイルスも，Human のレセプターをある程度は認識することがわかっている．そのため，この結果は，Avian のウイルスが連続変異を繰り返した後，直接 Human に感染したことを示していると考えられる．

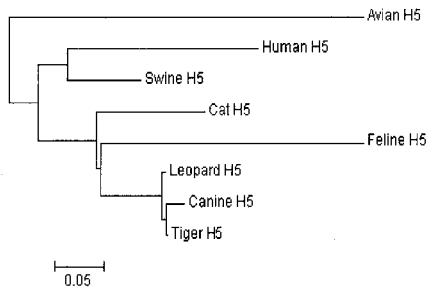

Fig.2 8 種の動物種から分離された H5 の進化系統樹

Fig.2 は，Avian（トリ科）,Canine（イヌ科），Cat（ネコ），Feline（ネコ科），Swine（ブタ）,Human（ヒト），Leopard（ヒョウ），Tiger（トラ）から分離された H5 のアミノ酸配列を用いて作成した NJ 法による進化系統樹である．Fig.2 から（Avian H5），（Swine H5，Human H5），（Canine H5，Cat H5，Feline H5，Leopard H5，Tiger H5）の 3 つのグループに分かれていることがわかる．まず，インフルエンザウイルスは野生の鳥類（Avian）のウイルスが起源とされていることを踏まえて（Avian H5）に注目すると，Avian H5 は早い段階で，Avian 以外に感染するタイプと Avian に感染するタイプに分かれたことが見て取れる [4]．次に（Swine H5,Human H5）を見ると，Swine H5 までの枝の長さが Human H5 までの枝の長さより短いことから，H5 が Avian→Swine→Human の順に変異していることがわかる．これは，先に述べた Swine の特異性を意味づけていると考えられる．最後に（Canine H5，Cat H5，Feline H5，Leopard H5，Tiger H5）を見ると，このグループは Feline と Canine から分離された H5 が分類されている．Feline と Canine には一般に A 型ウイルスは感染しないとされているが，WHO からの鳥インフルエンザの情報 (2)（2004 年 5 月 26 日）で A 型ウイルスの Feline への感染が報告されている．しかし，まだ Avian→Feline または Feline→Human という感染の可能性が低いようである．本研究結果においても，Avian，Feline，Human の H5 は別のグループに分類されている．しかし，Avian→Swine→Human のように，Avian→Feline→Human という感染が頻繁に起こるようになる可能性はある．また，遺伝的差異が小さいことから，Canine と Tiger に感染していた A 型ウイルスの H5 は類似のものであったと考えられる．

3.2 $RECD$ による分類

最も古い年代に採取されたアミノ酸配列 Influenza A virus (A/South Carolina/1/18 (H1N1)) (Host は Human) をインフルエンザウイルスの起源と仮定して基準配列 X とし，$RECD$ を求める際に対象となるその他の 3357 本の配列を $Y_i(i = 1, ..., 3357)$ として $RECD(X, Y_i)$ $(i = 1, ..., 3357)$ を計算した．最後に，その $RECD$ 値を小さい順に並べ替えてグラフ上にプロットした (Fig.3)．尚，Fig.3 の縦軸は $RECD$ 値を示し，横軸の No. は対象配列 Y_i を $RECD$ 値の小さい順に並べ替えた後に各 Y_i に付けた通し番号を意味する．$RECD$ による分類では，収集した全ての配列を使用した．

$RECD$ を用いることによって，HA 蛋白質を 5 つのグループに分類することが出来た．グラフの左から順にグループ 1，2，3，4，5 とおく．各グループの HA 蛋白質は以下の通りである．

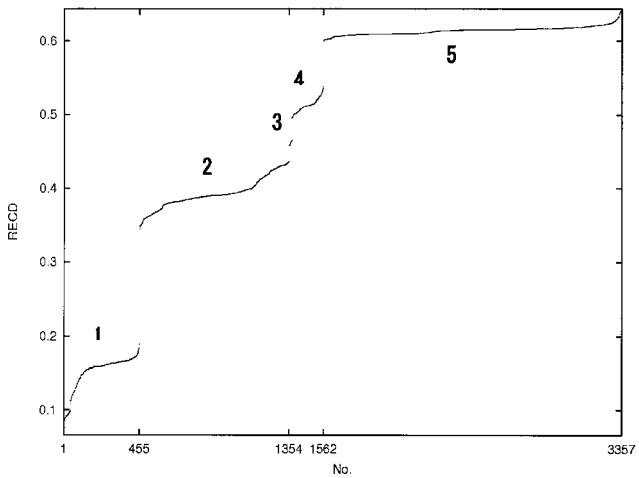

Fig.3 RECD による HA 蛋白質の分類結果

Table 1 インフルエンザ A 型ウイルス HA 蛋白質の分類

グループ	No.	RECD	HA の亜型
1	1〜455	0.0644〜0.1916	H1
2	456〜1354	0.3441〜0.4355	H2 H5 H6
3	1355〜1371	0.4585〜0.4666	H11
4	1372〜1562	0.4949〜0.5377	H8 H9 H12 H13 H16
5	1563〜3357	0.5994〜0.6427	H3 H4 H7 H10 H14 H15

抗原性による HA 蛋白質の分類では，16 種類に分類されていたが，$RECD$ による分類では 5 つのグループに分類することが出来た．この結果から，抗原性の分類ではわからなかった HA 蛋白質の類縁関係を見て取ることが出来る．また，$RECD$ は力学を考慮して定義されているので，この結果から，HA 蛋白質には 5 種類の変異力学が存在していることがわかる．

3.3 $RECD$ による HA 亜型別の変異解析

上記の配列，H1：456 本 (1918 年〜2005 年)，H2：72 本 (1957 年〜2005 年)，H3：1500 本 (1963 年〜2006 年)，H4：69 本 (1956 年〜2004 年)，H5：661 本 (1959 年〜2006 年) のアミノ酸配列を用いた (H6〜H16 のアミノ酸配列は登録本数が少ないため未使用)．それぞれ，最も古い年代に採取されたアミノ酸配列を，イン

フルエンザウイルスの起源と仮定して基準配列とし（Table 1），その基準配列と基準配列以外の対象配列との $RECD$ を計算した (Fig.4)．尚，Fig.6 の横軸は分離された年を意味し，時系列（分離された年順）となっている．縦軸は基準配列と対象配列との $RECD$ 値を示している．また，分離された年によっては複数本のアミノ酸配列が分離されているが，その場合は各年の $RECD$ 値の平均値を求めて上にプロットしている (Fig.4)．また，Fig.5〜Fig.9 は，H1〜H5 における $RECD$ の結果を用いて $Tan. = RECD(i+1) - RECD(i)$（尚，$i$ は，Fig.4 で示した年度に対する通し番号を表している．）を計算し，その結果をプロットしたものである．これにより，$RECD$ 値の変化の様子がより詳しくわかる．

Table 2 基準配列

HA の亜型	動物種	基準配列（Influenza A virus）
H1	Human	(A/South Carolina/1/18 (H1N1))
H2	Human	(A/Singapore/1/1957(H2N2))
H3	Equine	(A/equine/Uruguay/1/1963(H3N8))
H4	Avian	(A/duck/Czeckoslovakia/1956(N4N6))
H5	Avian	(A/chicken/Scotland/59(H5N1))

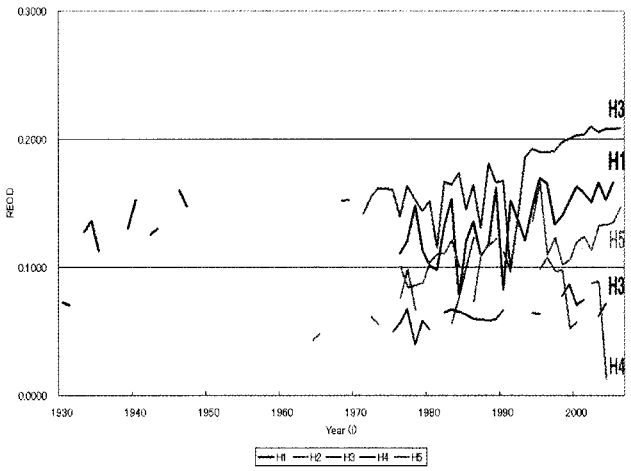

Fig.4 HA 亜型別の変異解析

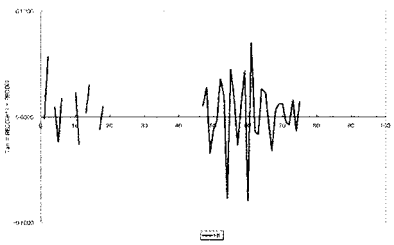

Fig.5 H1 の変異解析

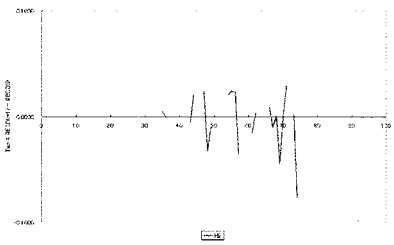

Fig.6 H2 の変異解析

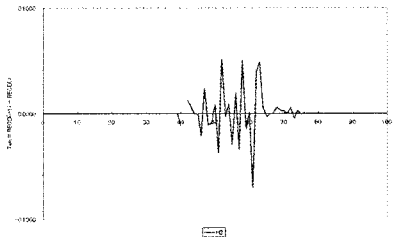

Fig.7 H3 の変異解析

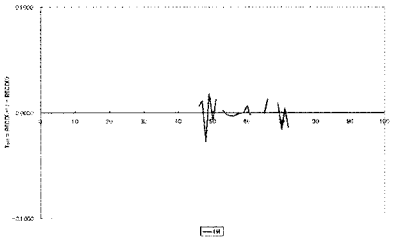

Fig.8 H4 の変異解析

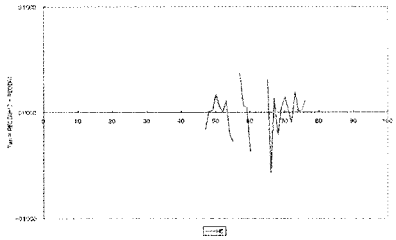

Fig.9 H5 の変異解析

H1，H3，H5 における *RECD* の変化は，どれも非常によく似ている（Fig.4）．特に，1980 年以降から *RECD* の軌道が似てきている．H1，H3，H5 は，最も流行しているウイルスの HA 亜型である．従って，この *RECD* の軌道は，流行しやすい HA 亜型の変異過程の特徴といえる．また，Fig.5〜Fig.9 で *RECD* の変化の様子を比較すると，Fig.5 と Fig.7 の軌道が非常に似ていることがわかる．H1 と H3 は 1918 年前後から流行している（1918 年以前から流行している可能性もある）HA 蛋白質の亜型である．H1 と H3 は，長い時間をかけて変異を繰り返すうちに，動物に感染しやすい（流行しやすい）変異を行うようになり，似たような変異の力学を保有するようになったと考えられる．このように，*RECD* を用いると数値的に生命の変異の特徴を捉えることが可能となる．

3.4 *RECD* による動物種別の変異解析

動物種として Avian，Equine，Human，Swine を選択し，Avian：59 本（1969 年〜2005 年），Equine：33 本（1971 年〜2003 年），Human：1355 本（1969 年〜2006 年），Swine：36 本（1980 年〜2006 年）を使用した．*RECD* と *Tan.* は HA 亜型別の変異解析と同じ手法を用いた．

Table 3 基準配列

HA の亜型	動物種	基準配列 (Influenza A virus)
H3	Avian	(A/duck/Ukraine/1963(H3N8))
	Equine	(A/equine/Miami/1/1963(H3N8))
	Human	(A/Hong Kong/1/68(H3N2))
	Swine	(A/swine/Colorado/1/1977(H3N2))

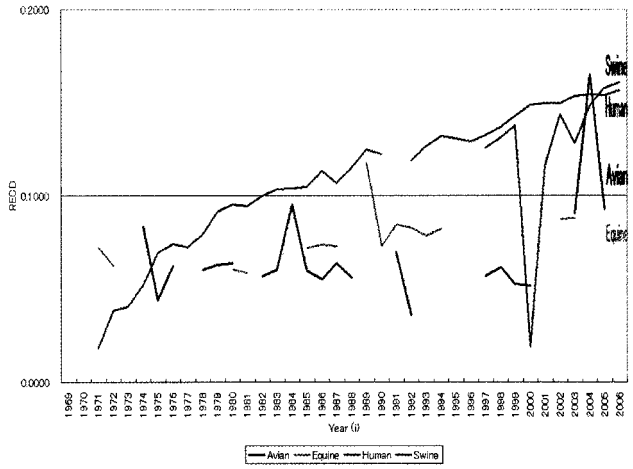

Fig.10 動物種別の変異解析

Fig.10 から，第 1 に，H3 における $RECD$ 値は，感染する動物種によって異なる変化をすることがわかる．従って，抗原性は同じであっても，感染する動物種によって H3 は異なる変異をする（異なる力学を保有している）と考えられる．第 2 に，Human の H3 における $RECD$ 値が単調に増加していることから，Human の H3 は変異（カオス）を蓄積しながら変異することがわかる．第 3 に，Human と Swine の $RECD$ 値の変化の軌道が 2004 年以降から似てきていることがわかる．従って，Human と Swine の H3 は，類似の力学によって変異すると考えられる．第 4 に，Avian と Equine の H3 は急に高い $RECD$ 値を取ることがわかる．このことから，Avian と Equine の H3 は，変異の過程で急激な変異を突然起こすと考えられる．第 5 に，Avian と Swine の H3 は，急に低い $RECD$ 値を取ることがわかる．これは，自然界に保存されていた過去に存在したウイルスが分離されたことを意味していると考えられる．

参考文献

[1] 梅垣壽春 大矢雅則 共著，確率論的エントロピー，情報科学講座 A・2・6，共立出版，(1983)
[2] 大矢雅則 著，情報進化論，岩波書店，(2005)
[3] 根井正利 著，五条掘孝・斉藤成也訳，分子進化遺伝学，培風館，(1990)
[4] 五藤秀雄，臨床医，vol.26 no.12，p.7-p.12，(2000)
[5] 中島節子，臨床医，vol.26 no.12，p.13-p.16，(2000)
[6] 松本慶蔵，臨床医，vol.26 no.12，p.86-p.87，(2000)
[7] M.Ohya (1991) Information dynamics and its application to optical communication processes, Springer Lecture Note in Physics, 378, 81-92.
[8] R.S.Ingarden, A.Kossakowski and M.Ohya (1997), Information Dynamics and Open Systems, Kluwer Academic Publishers.
[9] K.Sato and M.Ohya(2001) Analysis of the disease course of HIV-1 by entropic chaos degree, Amino Acids, 20, 155-162.
[10] Thompson, J. D., D. G. Higgins, and T. J. Gibson:CLUSTALW, Nucleic Acids Res. 22, 4673-4680 (1994)]