Vol.2010-MPS-81 No.26
Vol.2010-BIO-23 No.26
2010/12/17

情報学的理論を用いたインフルエンザウイルスの進化メカニズムの解明

山本眞吾[†]　　權 娟大[††]　　宮崎 智[††]

相互情報量を基礎として，アミノ酸配列中の共進化していると推定される残基を同定するための指標を考案し，インフルエンザウイルスの PB2 タンパク質を例に，共進化していると推定できる残基の推定と考察を行った．

Exploration of the evolutionally mechanism of the influenza virus using the information theory

Shingo Yamamoto[†] , Yeondae Kwon[††]
and Satoru Miyazaki[††]

We introduced a new measure based on mutual entropy in order to identify co-evolutional sites in influenza virus. By use of our measure, we analyzed more than 3,000 amino-acids sequences of PB2 protein and estimated the co-evolutional sites with 627th site of PB2 protein which is one of the most familiar sites associated with the pandemic.

1. 背景・目的

インフルエンザウイルスは過去に世界的大流行（パンデミック）を引き起こし，多数の死者を出してきた．これはインフルエンザウイルス遺伝子の非常に早い突然変異がウイルスのタンパク質構造を変化させ，ヒトが持つ免疫作用から逃れているためであると言われている．本研究はパンデミックを未然に防ぐために，突然変異の駆動力を探り,インフルエンザウイルスの進化メカニズムを解明することを目的としている．

現在多くの研究はインフルエンザウイルス表面にある HA（ヘマグルチニン）や NA（ノイラミニダーゼ）タンパク質に焦点を当てて研究されている．オセルタミビル（タミフル®）やザナミビル（リレンザ®）をはじめとする薬剤もノイラミニダーゼを阻害し，インフルエンザウイルスの増殖を抑えている．しかし，オセルタミビルではインフルエンザウイルスの突然変異によって耐性株が出現しているように，突然変異によるタンパク質の構造変化に対応できていないのが現状である．これまでに，これらのタンパク質遺伝子を合成しているインフルエンザウイルスの RNA ポリメラーゼのコピーエラーが突然変異の駆動力であると仮定している研究者も多い．しかし，RNA ポリメラーゼの立体構造データが十分であれば，突然変異を蓄積し易い残基の立体構造上の位置から，先の仮説を証明できる可能性もある．一方で，RNA ポリメラーゼの機能保持が欠かせないとすれば，パンデミックの引き金となる残基の突然変異と連動して，機能維持のための突然変異を起こす残基があり，そうした残基の中に，ミスコピー増大の原因となる突然変異残基が存在している可能性があろう．そこで，本研究では，２つの残基間の協調的な変異を検証するための測度としてエントロピー進化率を用い，共進化を起こしていると思われる残基の同定を試みた．

ヒトにとって強毒となる原因の１つに PB2 の 627 番残基のアミノ酸が GLu から Lys へ変異することが知られている[1]．627 番残基の変異を手掛かりに，エントロピー進化率の有用性の考察と 627 番残基と共進化したと推測できる残基群を同定したので報告する．

2. 準備

2.1 インフルエンザウイルス

インフルエンザウイルスはオルトミクソ科に属する（－）鎖 RNA ウイルスである．A～C 型に分類され，A 型には H5N1 のように表記される亜型が存在する．H5N1 の H

[†] 東京理科大学大学院　薬学研究科　薬学専攻
Graduate School of Pharmaceutical Sciences, Tokyo university of Science
[††]東京理科大学　薬学部　生命創薬科学科
Faculty of Pharmaceutical Sciences, Tokyo University of Science

Vol.2010-MPS-81 No.26
Vol.2010-BIO-23 No.26
2010/12/17

とNはウイルス表面タンパク質のヘマグルチニンとノイラミニダーゼに由来し，それらの抗原性によってH1～H16，N1～N9に分けられる．A型はヒト，トリ，ブタ，ウマなどに感染し，B，C型はヒトのみに感染する．

インフルエンザウイルスゲノムは八つの分節（segment）を持つ．一つの分節から1つあるいは2つのタンパク質が合成される．合成されるタンパク質はA型～C型それぞれ異なる．表1に分節と合成されるタンパク質の関係をまとめた．PB1，PB2，PAは三量体を形成し，RNAポリメラーゼとして働いている．またHA，NAは宿主細胞の吸着と遊離に関与する酵素である．

表1 インフルエンザウイルスの分節と合成されるタンパク質

segment	IVA	IVB	IVC
1	PB2	PB1	PB2
2	PB1 PB1-F2	PB2	PB1
3	PA	PA	PA
4	HA	HA	HE
5	NP	NP	NP
6	NA	NA	M1
		NB	CM2
7	M1	M1	NS1
	M2	BM2	NS2
8	NS1	NS1	
	NS2	NS2	

インフルエンザウイルスは過去にパンデミックを引き起こしている．これまでの大流行について表2にまとめた．

表2 過去に起きたパンデミック

年	名称・流行地域	亜型
1918	スペイン風邪	H1N1
1933	ロンドン風邪	H1N1
1957	アジア風邪	H2N2
1968	香港風邪	H3N2
1977	ソ連風邪	H3N2
1997	香港	H5N1
2001	中近東，欧米，北アフリカ	H1N2

2.2 インフルエンザウイルスの配列情報の取得

国際塩基配列データベースDDBJ（DNA databank of Japan）におけるGIB-V（Genome Information Broker for Viruses）を使用した．Genome Listから Orthomyxoviridae科のInfluenza のデータを flat file形式で取得した．

flat file のデータから ACCESSION番号，亜型，種，地域，登録番号，年，分節，年月日，CDS配列，塩基配列を perl言語により csv形式で抽出した（表3を参照）．

表3 抽出したデータと内容

抽出したデータ	データの説明
ACCESSION番号	DDBJアクセッション番号
亜型	配列の得られた生物，ウイルスなどの血清学的タイプ
種	配列の得られた生物の学名
地域	配列サンプルを得た地域
登録番号	インフルエンザウイルスの分離された順番を表す番号
年	配列サンプルを得た年
分節	配列の得られたウイルスのセグメント
年月日	標本を採取した日付
塩基配列	塩基配列
CDS配列	タンパク質のアミノ酸をコードする配列

こうして国際塩基配列データベースで公開されているゲノム配列を中心に，1967年から2007年までの40年間について，トリ，ヒト，ブタで報告されたウイルス全ての配列を収集し独自のデータベースに格納した．

3. 情報理論を用いたウイルス配列の解析

本研究では，PB2 タンパク質を例にその変異の特徴付けとして，変異の入り方，すなわち,配列に蓄積された置換に共通の法則を見出すことを試みる．そのために，毒性を増すと言われている627番残基の変異に着目して解析を行った．759個の残基のうち，627 番残基の置換に伴う別の残基でのアミノ酸置換は少ないと考えられる．一般に，各残基で生起するアミノ酸残基の種類と割合は残基ごとに異なっている．これらの前提を考慮すれば，直接の対応関係は見られない場合，置換の対応に規則性がある残基の検出を行える測度が必要である．そのためには，これまでの遺伝学的距離に基づくような相同性あるいは同一性を超えて，対応する変異の関係性を考慮できる必要がある．そこで本研究ではアミノ酸配列解析にはエントロピー進

Vol.2010-MPS-81 No.26
Vol.2010-BIO-23 No.26
2010/12/17

化率という情報理論を用いた[2]．以下，計算過程において必要なシャノンエントロピーと相互情報量について具体的に説明する．

3.1 シャノンエントロピーの計算方法

シャノンエントロピー（Shannon Entropy，以下 SE とする）は乱雑さを表す尺度であり，配列におけるアミノ酸の出現の偏りを知ることができる．シャノンエントロピーは情報分野における通信の技術などに役に立っている理論であり，情報分野以外にも広く応用されている．アミノ酸が均等に出現しているとき最もシャノンエントロピーの値が高く，アミノ酸の偏りが大きくなるほどその値は小さくなっていく（0≦SE≦2）．シャノンエントロピーを求める式は以下のように表される．P(x)はアミノ酸の出現確率を表しており，アミノ酸の数で和を取る．

$$SE(P) = -\sum P(X)\log P(X)$$

例えば，図 2 のようなある配列 A と配列 B が与えられたとき，A と B の SE は，

$$SE(\text{配列}A) = \left(-\frac{3}{10}\log\frac{3}{10}\right) + \left(-\frac{3}{10}\log\frac{3}{10}\right) + \left(-\frac{4}{10}\log\frac{4}{10}\right) = 1.570951$$

$$SE(\text{配列}B) = \left(-\frac{6}{10}\log\frac{6}{10}\right) + \left(-\frac{4}{10}\log\frac{4}{10}\right) = 0.970951$$

と，求められる．

3.2 相互情報量の計算方法

相互情報量（Mutual Information，以下 MI とする）はある２つの情報源 A と B がどれだけ関連しているのかを表す尺度である．A と B のアミノ酸の出現に関連がない場合，MI は 0 になる．逆に，A の配列が分かれば B の配列も分かる場合，2 の値を取る（0≦MI≦2）. 相互情報量を求める式は以下のように表される. P(X,Y)は配列 A と B において二つのアミノ酸が同時に出現する確率を表し，P(X)，P(Y)はそれぞれ配列 A，配列 B においてアミノ酸が出現する確率を表しており，二つの配列のアミノ酸の数で和を取る．

$$MI(A,B) = \sum_{X,Y} p(X,Y)\log\frac{p(X,Y)}{p(X)p(Y)}$$

例えば，図 2 の配列 A と配列 B の相互情報量 MI は，

$$MI(\text{配列}A,\text{配列B}) = \frac{3}{10}\log\left(\frac{\frac{3}{10}}{\frac{3}{10}\times\frac{6}{10}}\right) + \frac{3}{10}\log\left(\frac{\frac{3}{10}}{\frac{3}{10}\times\frac{6}{10}}\right) + \frac{4}{10}\log\left(\frac{\frac{4}{10}}{\frac{4}{10}\times\frac{4}{10}}\right) = 0.970951$$

と，求められる．

3.3 エントロピー進化率の計算方法

エントロピー進化率（Entropy Evolutional Rate，以下 EER とする）は MI を正規化した値である．MI は SE の大きさに依存するため，MI の比較は相対的なものとなってしまう．そこで，EER によって MI を正規化することで，二つの情報源を正しく評価できる．EER は以下の式により定義される．

$$EER(AB) = \frac{MI(A,B)}{SE(A)+SE(B)-MI(A,B)}$$

例えば，図 2 の配列 A と B の EER は，

$$EER(AB) = \frac{0.970951}{1.570951+0.970951-0.970951} = 0.618066$$

と，求められる．この値が大きいほど関連性が高いと言える．EER は 0≦EER≦1 の値を取る．

図 2　配列 A と配列 B での SE，MI，EER の計算例

4. 実験

変異によって強毒性を発揮する 627 番残基とその他の残基との関連性を測定するために，まず PB2 タンパク質のアミノ酸配列を 1976 年から 2007 年までを抽出し，年度ごとにマルチプルアラインメントを行った. 次に, マルチプルアラインメントした PB2 タンパク質のアミノ酸配列において，627 番残基とそれ以外の残基との n-1 通り（n は配列長）の EER を計算した（図 3）．

Vol.2010-MPS-81 No.26
Vol.2010-BIO-23 No.26
2010/12/17

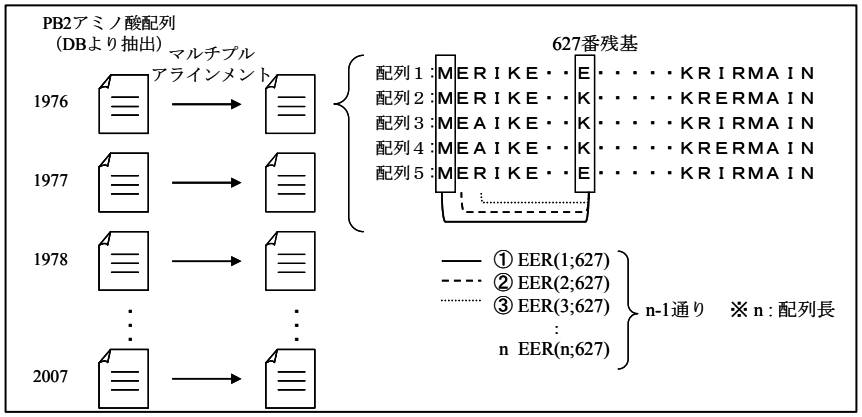

図 3　627 番残基とその他の残基の関連性測定の計算方法

表 4　A~C 型のレコード数

type	record
IVA	26413
IVB	1502
IVC	13
total	27928

表 5　種別 A 型インフルエンザウイルス のレコード数

A 型	environment	equine	human	mouse	swine	avian	blank	total
1	56	36	2315	2	105	750	1	3265
2	57	36	2320	2	105	778	1	3299
3	56	36	2314	2	105	761	1	3275
4	55	36	2316	2	105	746	0	3260
5	56	37	2315	2	105	792	1	3308
6	56	37	2364	2	105	742	0	3306
7	56	37	2368	2	105	787	1	3356
8	56	37	2354	2	105	789	1	3344

表 6　亜型，地域，年のレコード数

亜型		地域		年	
subtype	record	place	record	year	record
H3N2	12714	New York	7583	2007	3001
H1N1	6647	Canterbury	2496	2005	2599
H5N1	958	Ohio	1152	2003	2255
H7N2	658	Memphis	1071	2000	2109
H3N8	636	ALB	939	2004	2027
H4N6	367	Waikato	816	2002	1861
H7N3	358	Western Australia	699	2001	1467
H6N2	340	Wellington	664	1999	1286
H7N1	298	Kentucky	632	2006	852
H2N2	247	Italy	585	1996	701
:	:	:	:	:	:
total	26413	total	26413	total	26413

5. 結果・考察

5.1 データベースの構築

現在，インフルエンザウイルスに関するデータベースは存在しない．そこで，DDBJ に登録されている 27,928 件のインフルエンザウイルスの情報を取得し，付加情報を付与した独自のインフルエンザウイルスデータベースを構築した．表 4 に A～C 型のレコード数をまとめた．この表からデータの 95%近くが A 型であることが分かる．表 5 に種別 A 型インフルエンザウイルスのレコード数をまとめた．表 6 に亜型，地域，年のレコード数のトップ 10 を示す．亜型（HxNx）はヒトに感染する H3N2 や H1N1 が 70%以上を占めていて，採取された地域はニューヨークが最も多かった．また，採取された年に関しては，年が新しいほどデータ数が多くなっていく傾向があった．

5.2 EER による解析結果

本研究では，数学的測度として二つの配列間の「変異の関係」を記述できる MI を選定した．EER は，MI を正規化して，その値域を 0 と 1 の間に取るようにした．ただし，その配列自身にばらつきがなければ，配列自身の情報量が 0 であるとみなされ

Vol.2010-MPS-81 No.26
Vol.2010-BIO-23 No.26
2010/12/17

るので，二つの配列間の EER も自動的に 0 となる．すなわち，EER は配列対象となる残基に変異がない場合は意味をなさない．

インフルエンザウイルスの 627 番残基 Glu は，パンデミックが起きたとしても，すべて Lys とはならず，ある程度の割合で Glu と Lys が混ざっている．本研究では，アミノ酸変異が起こる残基に着目してその特性を明らかすることを目的としているので，変異が生じない残基は今後自動的に排除されることが望ましい．これらの側面からも情報量を基にした測度を用いた解析手法の提案は有用であると考えられる．

本研究では，EER を用いることによって，PB2 タンパク質の 627 番残基の Lys と Glu の割合の変異に着目し，変異の仕方に相関のある 627 番残基と共進化していると思われる残基の推定を試みた．

解析に用いたパンデミックに対応した年，1977，1997，2001 年の 627 番残基の Lys の割合は，それぞれ 54%，63%，52%であった．パンデミックの年において，627 番残基の変異と対応して EER の値が 1，すなわち，627 番残基の変異と同じようなルールで変異した残基が各年においていくつかあることが分かった．これらの残基は，627 番の変異と対応した変異が起こっているので，627 番残基と何らかの生化学的特性がある可能性がある．さらに，1976 年から 2007 年のデータにおいて，1977，1997，2001 年以外の年について，627 番残基との関連性を考察すると，パンデミックのあるなしに関わらず，常に同様の変化をしている残基がいくつかあることが分かった（表 7）．これらの残基は，627 番残基が Glu から Lys に変異すると同様に，Met から Leu への変異というような，その生起確率と 627 番残基とのアミノ酸の対応を維持するような共進化的な変異が存在していた．こうした残基ではパンデミックに依存せず，共進化の関係が継続していた（表 8）．さらに，30 年間のデータを時系列でみたときに，エントロピー進化率が 0.1 以下，すなわち，627 番残基とはほぼ独立であるか，または変異が認められない残基が全体の約 9 割に達していることが分かった．

本研究では，インフルエンザウイルス配列の薬剤耐性獲得の駆動力を推定することを大きな目標としている．薬剤耐性をもつための条件として，配列変異の大きさが挙げられており，この駆動力として RNA ポリメラーゼである PB1，PB2，PA タンパク質の複合体のコピーエラーを挙げる研究者も少なくない．この仮定が正しければ，PB2 においてもコピーエラーの原因となる変異がパンデミックに特異的に起こっていると考えられる．こうした変異はある残基に特異的に起こっている可能性もあるが，高次構造的な制約があるとすると，いくつかの残基に連鎖的に起こっている可能性が高い．本研究では，627 番残基と他残基間の関連性のみに着目したが，全ての残基の組み合わせを網羅的に数量化するとともに，残基ごとの時系列的な変化を調べた．オーソログ配列の比較とは違い，各残基に生起するアミノ酸種類は限定的であると考えられる．そのため遺伝学的距離だけでなく，確率分布を考慮した「情報量」という観点から時系列変化を数量化した解析が有効となる可能性が高いと考えられる．

表 7　1976 年~2007 年間の EER の平均値の高い残基

残基番号	平均値
475	0.903255886
199	0.901197342
368	0.849402545
661	0.771319171
64	0.757413304
9	0.750995467
44	0.677192377
271	0.667853389
567	0.653713134
702	0.652860811

表 8　1976~2007 年における 475 番残基の EER の推移

年	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986
EER	1	1	1	1	1	1	1	0.72	1	1	1
K*	65	54	53	13	49	32	27	68	34	29	26

年	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
EER	0.73	0.83	0	1	1	1	1	1	1	1	1
K	26	28	6	19	27	10	86	93	90	93	63

年	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	ave
EER	1	1	0.86	1	1	-	0.86	0.87	0.14	1	0.903
K	77	76	86	91	82	88	79	52	66	99	55.84

（K*：K の出現確率）

Vol.2010-MPS-81 No.26
Vol.2010-BIO-23 No.26
2010/12/17

6. 参考文献

[1] Shinya K., Hamm S., Hatta M., Ito H., Ito T., and Kawaoka Y.: PB2 amino acid at position 627 affects replicative efficiency, but not cell tropism, of Hong Kong H5N1 influenza A viruses in mice, Virology, Vol.320, No.2, pp.58-266 (2004).
[2] Ohya M.: Information theoretical treatment of genes, Trans. IEICE, E72, No.5, pp.556-560 (1989).