様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １１ 日現在

機関番号：１６４０１
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５７０２４９

研究課題名（和文）タンパク質にかかる多様化圧の時空間集積性および適応コストと補償的変異のベイズ評価

研究課題名（英文）Spatial distribution of selection pressure on a protein deriving its adaptation to the environment based on the hierarchical Bayesian model

研究代表者
渡部　輝明（WATABE, Teruaki）

高知大学・教育研究部医療学系・講師

研究者番号：９０３２５４１５

交付決定額（研究期間全体）：（直接経費）　　4,300,000円

研究成果の概要（和文）：タンパク質の環境適応はアミノ酸配列置換によってもたらされるが、それは遺伝子上で起こった突然変異が淘汰された結果である。そのため突然変異の淘汰は環境に特有の選択圧のもとで行われると考えられる。つまりタンパク質適応進化について理解するには、選択圧の空間的な分布を検出することが重要となる。我々は階層ベイズモデルを通して、タンパク質表面における選択圧の空間分布を検出する方法を開発した。事前分布にポッツ模型を採用し、選択圧の空間集積性の強さと広さを決める超パラメータは周辺尤度を最大化することで決定可能である。この方法をインフルエンザウイルスのヘマグルチニンタンパク質に適用し、選択圧の空間分布を検出した。

研究成果の概要（英文）：Proteins adapt to novel environments and/or gains functions by substitution in amino-acid sequences. Therefore, mutations in protein-coding genes occur under the selection pressure, the strength and character of which may vary among the regions of the protein. Thus, the spatial distribution of selection pressure provides information on the adaptive evolution of the protein. To detect the distribution, we developed a hierarchical Bayesian model. The Potts model describes the prior distribution of spatial aggregation of selection pressure. The hyper-parameters that define the strength and range of spatial clustering are estimated by maximizing the marginal likelihood. We applied the method to historical data on the influenza hemagglutinin protein, comparing the estimated spatial distribution to that of antigenic sites A-E. The amino-acid residues with higher substitution-rate ratios, representing diversifying selection pressure, overlapped the antigenic sites.

研究分野：分子進化生物学

キーワード：分子進化　タンパク質多様化圧

１．研究開始当初の背景

タンパク質はアミノ酸配列を置換することで機能を獲得又は変化させて環境に適応していくが、遺伝子上で起こった突然変異は環境に特有の選択圧のもとで淘汰されていく。そのため選択圧はタンパク質表面の機能領域か否かに依存すると同時に、進化の過程で環境にも依存していくものと考えられる。タンパク質上の同じ空間的位置においても進化過程という時間的な位置によって適応的変異であるか中立的変異であるかが異なってくる（Ref. 1, 2）。選択圧の時空間的揺らぎを定量的にとらえることで、タンパク質適応進化機構を探ることが出来る。

タンパク質の配列変異にかかる選択圧は、塩基配列における非同義置換と同義置換の速度比（$\omega = d_N/d_S$）で表現され、配列の突然変異率へ取り込むことでモデル化される。これまでは一次配列情報のみを用いた選択圧検出の解析が主流であったが、情報量の不足から選択圧の一次配列上での分布のみが検出されてきた（Ref. 3）。またタンパク質立体構造の情報を取り入れた空間分布の検出（Ref. 4）も試みられているが、局所尤度法の考えを用いていることから内包するパラ－メータの決定が難しく、また進化過程での時間分布を検出することは困難と言える。

２．研究の目的

本研究では階層ベイズモデルを通してタンパク質表面における選択圧の空間分布を検出する方法を開発し、インフルエンザ（Flu）ウイルスのヘマグルチニン（HA）タンパク質に適用することで、選択圧の空間分布を検出する。HA タンパク質は、Flu ウイルスの主要な抗原となるタンパク質であり、そのアミノ酸配列が緩やかに変異することで Flu ウイルスがヒト宿主集団に感染し続けることを可能にしている。しかしヒト宿主集団の交差免疫の作用により、Flu ウイルスは遺伝的な多型を残すことが出来ず、一次元的な自由度で遺伝情報を伝え残している。そして抗原性に断続的に生じている大きな変異によって幾つかのクラスターに分割される。このクラスター構造の発生機序は未だ解明されておらず、配列変異に生じた選択圧の時空間分布を明らかにし、その選択により抗原性の変異がどのような物理化学的変化によってもたらされたのかを明らかにすることでクラスター構造発生機序を解明することが目的である。

３．研究の方法

(1)周辺尤度

タンパク質立体構造（$\boldsymbol{X}$）が与えられたときの遺伝子配列の条件付き確率は以下で与えられる：

$$Z = P\left(\left\{\boldsymbol{A}^{(i)}\right\}\middle|\boldsymbol{X}\right)$$

ここで$\boldsymbol{A}^{(i)}$は配列長がLのコドン配列を表しており、$i = 1, \cdots, N$の範囲にわたる。配列変異にかかる選択圧は、塩基配列における非同義置換と同義置換の速度比（$\omega = d_N/d_S$）で表現され、以下の様に明示的に導入される：

$$Z = \int P\left(\left\{\boldsymbol{A}^{(i)}\right\}\middle|\boldsymbol{\omega}, \boldsymbol{X}\right)P(\boldsymbol{\omega}|\boldsymbol{X})d\boldsymbol{\omega}$$

各アミノ酸残基における置換速度比はベクトル表記されており、$\boldsymbol{\omega} = (\omega_1, \cdots, \omega_L)$、期待値として求めることが出来る：

$$\langle\omega_k\rangle = \frac{1}{Z}\int \omega_k P\left(\left\{\boldsymbol{A}^{(i)}\right\}\middle|\boldsymbol{\omega}, \boldsymbol{X}\right)P(\boldsymbol{\omega}|\boldsymbol{X})d\boldsymbol{\omega}$$

(2)熱力学的積分

事前分布$P(\boldsymbol{\omega}|\boldsymbol{X})$が規格化されていない場合、周辺尤度の表記には事前分布の積分を分母に持つ必要がある：

$$Z = \frac{\int P\left(\left\{\boldsymbol{A}^{(i)}\right\}\middle|\boldsymbol{\omega}, \boldsymbol{X}\right)P(\boldsymbol{\omega}|\boldsymbol{X})d\boldsymbol{\omega}}{\int P(\boldsymbol{\omega}|\boldsymbol{X})d\boldsymbol{\omega}}$$

しかし一般的にパラメータの全空間に渡る積分は現実には困難を伴う。そこで周辺尤度の対数をとり、右辺が分子の対数と分母の対数の差であることを形式的にパラメータβの導入により表現する：

$$\ln Z$$

$$= \int_0^1 \frac{d}{d\beta}\left\{\ln \int \left\{P\left(\left\{\boldsymbol{A}^{(i)}\right\}\middle|\boldsymbol{\omega}, \boldsymbol{X}\right)\right\}^{\beta} P(\boldsymbol{\omega}|\boldsymbol{X})d\boldsymbol{\omega}\right\}d\beta$$

この後はパラメータβによる微分を遂行し、パラメータ$\boldsymbol{\omega}$の全空間に渡る積分を回避する形式を得る。パラメータβにより特徴付けられる系における期待値を計算することで周辺尤度の対数を得るのである：

$$\ln Z = \int_0^1 E_\beta\left[\ln P\left(\left\{\boldsymbol{A}^{(i)}\right\}\middle|\boldsymbol{\omega}, \boldsymbol{X}\right)\right]d\beta$$

ここで$E_\beta[\cdots]$は以下の密度関数を考慮した期待値を意味する：

$$d_\beta(\boldsymbol{\omega}) = \frac{\left\{P\left(\left\{\boldsymbol{A}^{(i)}\right\}\middle|\boldsymbol{\omega}, \boldsymbol{X}\right)\right\}^{\beta} P(\boldsymbol{\omega}|\boldsymbol{X})}{\int\{P(\{\boldsymbol{A}^{(i)}\}|\boldsymbol{\omega}, \boldsymbol{X})\}^{\beta} P(\boldsymbol{\omega}|\boldsymbol{X})d\boldsymbol{\omega}}$$

(3)ポッツ模型を用いたギブスサンプリング

期待値$E_\beta[\cdots]$を計算するためにギブスサンプリングを用いる。置換速度比の事前分布はω_kをK個の状態（$s_k = 1, \cdots, K$）に階級化したポッツ模型によりモデル化する：

$$P(\boldsymbol{s}|\boldsymbol{X}) = \exp\left\{\lambda \sum_{l>k} \boldsymbol{Q}_{s_k s_l} \exp(-\alpha r_{kl})\right\}$$

この模型は、格子上スピンの相互作用を記述するイジング模型を拡張し、3 次元空間に分布する素子間の相互作用をそれらの空間距離を考慮して扱えるようにしたものである。タンパク質の 3 次元構造はアミノ酸残基のα炭素原子間の空間距離r_{kl}のみを用いる。$\boldsymbol{Q}$はK行K列行列であり、対角要素が$(K-1)/K$で非対角要素が$-1/K$である。この事前分布では近距離にあるアミノ酸残基の組は遠距離

にある組より高い相関を持つことが保証されている。

アミノ酸配列の進化過程で各アミノ酸残基での置換は、タンパク質の 3 次元構造においてその残基が位置する部位の担う機能により影響を受けると考えられる。ここではそれらの影響を置換速度比としてのみ表現し、配列が与えられた時のωの尤度を各残基での尤度の積として表す：

$$P\left(\left\{\boldsymbol{A}^{(i)}\right\}\middle|\boldsymbol{\omega},\boldsymbol{X}\right)=\prod_{k=1}^{L}P(\boldsymbol{a}_k|\omega_k)\,,$$

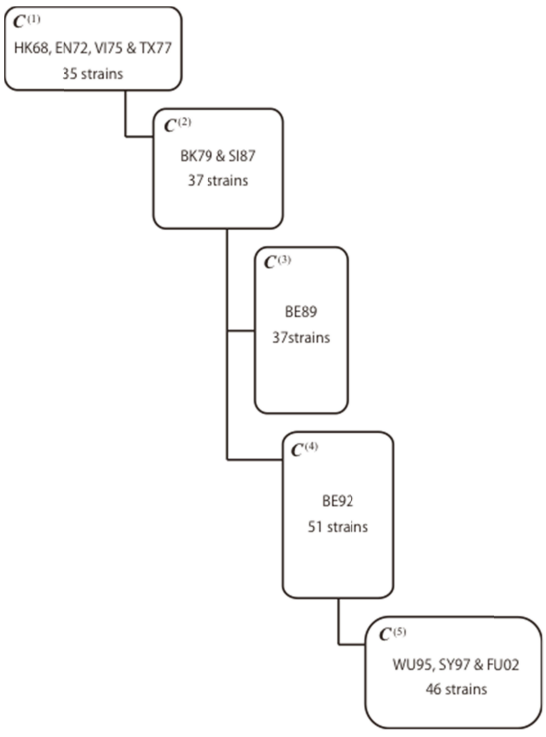
ここで$\boldsymbol{a}_k$はN本のアミノ酸配列のk番目のアミノ酸を要素とするベクトルである：

$$\boldsymbol{a}_k=\left(a_k^{(1)},\cdots,a_k^{(N)}\right)$$

各階級（$s_k=1,\cdots,K$）でのω_kの値は、サンプリングの 1 ステップ毎に尤度を最大にするように更新される。10 万ステップのギブスサンプリングを行い、開始後 2 万ステップを安定期に入るまでのバーンインとして扱う。安定期の 8 万ステップを用いて$\langle\omega_k\rangle$の推定を行う。

(4)抗原性の断続変異

Flu ウイルス（H3N2 型）の 1968 年から 2003 年までの系統関係を断続的な抗原性の変異によって 11 のクラスターに分けることが出来る（下図、Ref. 5）。



11 のクラスターのうち、幾つかは内包する配列数が比較的少ないため、それらをまとめて 5 つのクラスター（$\boldsymbol{C}^{(1)}$～$\boldsymbol{C}^{(5)}$）にした。図は系統関係の年代順を大まかに表している。それぞれのクラスターは 35 から 51 本の配列を有している。これら 5 つのクラスターのうち、$\boldsymbol{C}^{(3)}$は系統関係上、後に続くクラスターがない“絶滅”したクラスターである。

異なるクラスターでは異なる選択圧分布を持つことを想定して、新たな自由度$\boldsymbol{\varepsilon}$を導入する：

$$Z=\int P\left(\left\{\boldsymbol{C}^{(i)}\right\}\middle|\boldsymbol{\varepsilon},\boldsymbol{\omega},\boldsymbol{X}\right)P(\boldsymbol{\varepsilon},\boldsymbol{\omega}|\boldsymbol{X})d\boldsymbol{\varepsilon}d\boldsymbol{\omega}$$

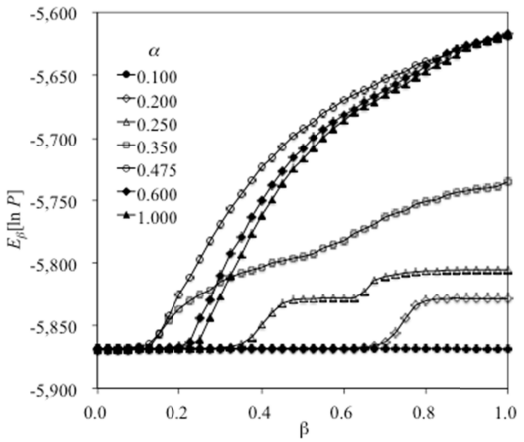
ここで$\boldsymbol{C}^{(i)}$はクラスターを表す。新たな自由度$\boldsymbol{\varepsilon}$は背景となる分布$\boldsymbol{\omega}$からの逸脱を表し、クラスター毎に指定される。

4．研究成果

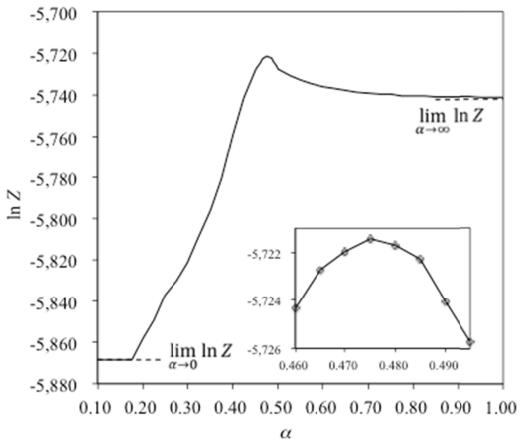
暫くはクラスターの導入以前における研究成果について記述する。

(1)超パラメータの推定

数値積分により熱力学的積分を遂行して、周辺尤度を最大化する超パラメータ(λ,α)を求める。パラメータβの範囲$0\le\beta\le1$を 40 の区画に分割し、各値でギブスサンプリングを行った。下図に期待値$E_\beta[\ln P]$のβ依存を$K=3$及び$\lambda=3.0$の場合で示した。

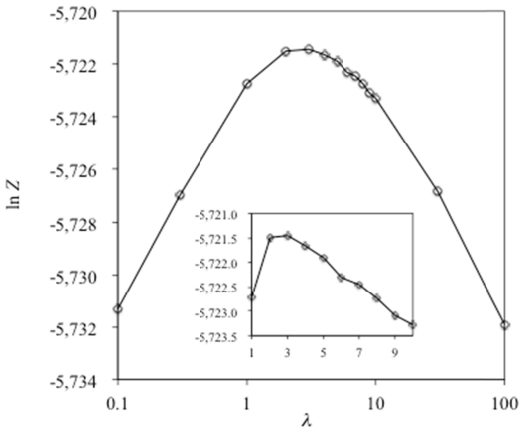


期待値$E_\beta[\ln P]$のβ依存は非常に滑らかであり、これらの数値積分により対数尤度$\ln Z$のα依存が得られる（下図）。



$K=3$及び$\lambda=3.0$の場合では$\alpha=0.475$において最大尤度を得た（$\ln Z=-5721.44$）。下図には$0.1\le\lambda\le100$の範囲で同様にして求めた各λにおける最大尤度を示した。この結果から$K=3$では$(\lambda,\alpha)=(3.0,0.475)$において最大周辺尤度を得ることが判る。これを

$M^{(K3)}$モデルと呼ぶことにする。また、$K = 10$では$(\lambda, \alpha) = (4.0, 0.450)$において最大周辺尤度を得ることが判り、$M^{(K10)}$モデルとする。

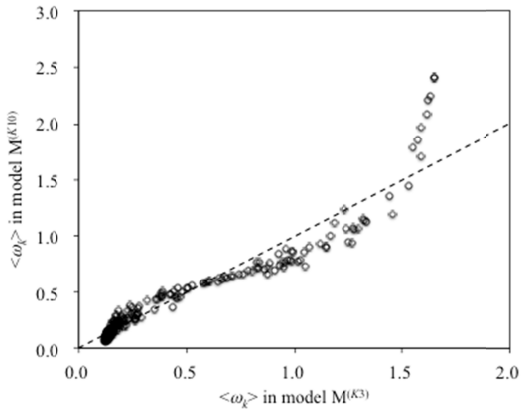


(2) $\langle\omega_k\rangle$の推定

$M^{(K3)}$モデルにおける各階級値$\hat{\omega}^{(s)}$をその利用率とともに下の表に示す。

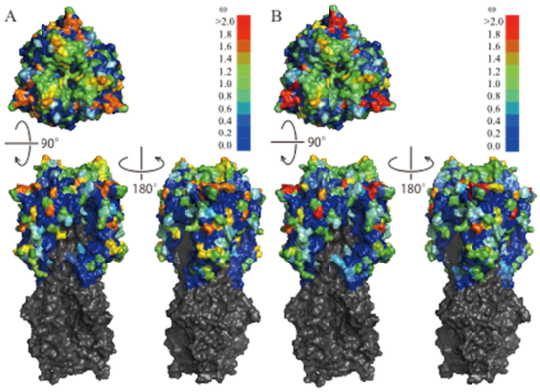
s	1	2	3
$\hat{\omega}^{(s)}$	0.120	0.766	1.648
利用率	0.714	0.183	0.103

これらの値は$\beta = 1$に設定したギブスサンプリングから得られるものと同じである。階級 1 での$\hat{\omega}^{(s)}$は変異の浄化をもたらす値であり、浄化圧が 70%以上のコドン座位にかかっていたことを示している。一方で階級 3 では$\hat{\omega}^{(s)}$は多様化を促す値となっており、10%程度のコドン座位で多様化圧がかかっていたと考えられる。$M^{(K10)}$モデルにおいても同様の結果を得た。$\hat{\omega}^{(s)} < 0.5$の値を階級 1 から 4 の 4 階級が取り、70%強のコドン座位が強い浄化圧を受けていたことが示された。多様化をもたらすのは階級 10 のみであり、その階級値は$\hat{\omega}^{(s)} = 2.410$と高いものであったが、利用率は低く 3.6%であった。これらから得られる$\langle\omega_k\rangle$は以下の様になっている。

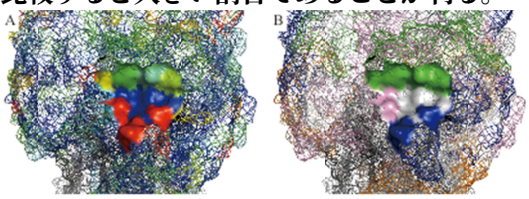


この図は$M^{(K3)}$モデルと$M^{(K10)}$モデルで推定された$\langle\omega_k\rangle$の対応を示している。概ね一致しているが$M^{(K10)}$モデルはより高い値を実現している。

次にタンパク質上での分布を示す。下図 A では$M^{(K3)}$モデルの結果を示し、図 B では$M^{(K10)}$モデルの結果を示している。

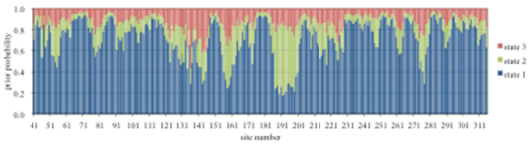


更に抗原領域との比較を行った。高いd_N/d_S比（$\langle\omega_k\rangle > 1.0$）を示すアミノ酸残基は抗原領域と重なっており、抗原領域においてその様な残基は$M^{(K10)}$モデルで 14.5%を占めていた。これはタンパク質全体での割合 7.6%と比較すると大きい割合であることが判る。



上図では宿主細胞受容体との結合領域を強調表示している。図 B に示されているように抗原領域 A（青）、B（緑）と D（桃）が結合領域に重なりを有している。結合領域に限るとその 40.0%で高いd_N/d_S比（$\langle\omega_k\rangle > 1.0$）を示した。

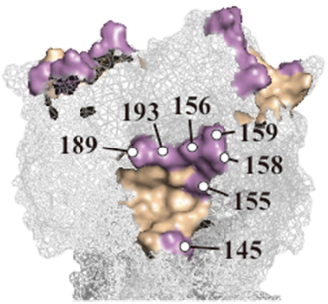
下図に$M^{(K3)}$モデルでの事前確率を示した。



アミノ酸残基（コドン座位）毎に各階級の利用率を示しており、明らかな部位依存が認められる。

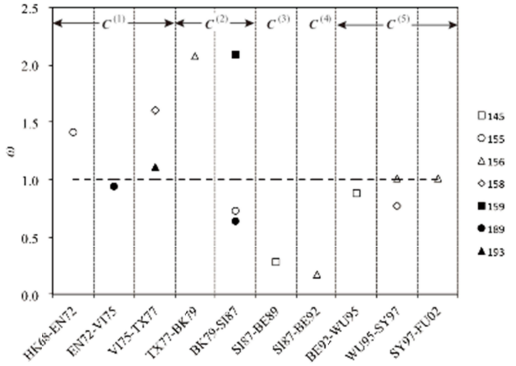
ここからはクラスターを導入した後の研究成果について記述する。

(3)受容体結合領域周辺のアミノ酸残基



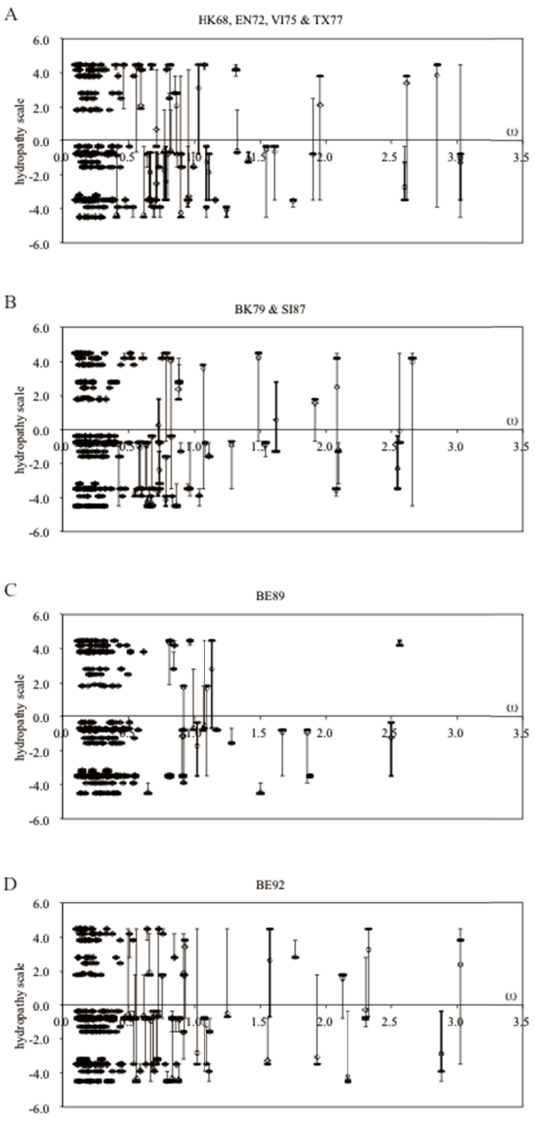
上図で示した受容体結合領域周辺の 7 つのアミノ酸残基で起きた 1 アミノ酸置換が、主要な抗原性の変異をもたらした置換である

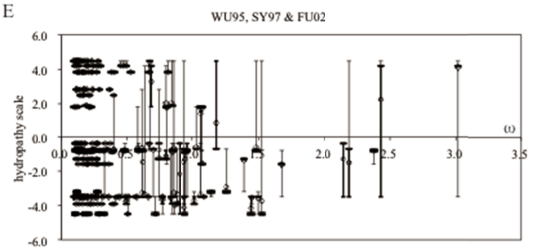
ことが示されている（Ref. 6）。各クラスターにおける、それら 7 つの残基でのd_N/d_S比を下図に示した。



一般に受容体結合領域のように機能に直結した部位での変異は有害な場合が多く、d_N/d_S比は低いことが想定される。実際にも上図のほとんどの場合で中立か浄化を示している。しかし$\boldsymbol{C}^{(1)}$と$\boldsymbol{C}^{(2)}$では、対応するアミノ酸残基において多様化もたらす置換が起きていたことが示された。

(4)多様化圧と物理化学的変異





上図は、HA タンパク質の各アミノ酸残基で起きた置換によって、どの程度の疎水性の変化が生じたかをクラスター毎に表したものであり、d_N/d_S比との関連と共に示している。一般に多様化圧の下での置換は、そのタンパク質の物理化学的な性質を変化させて、何らかの機能的な変異をもたらすと考えられる。しかし上図 C が示しているように、$\boldsymbol{C}^{(3)}$（BE89）での高いd_N/d_S比（$\langle\omega_k\rangle > 1.0$）を伴う置換が、疎水性の尺度の符号を変えることなく起きていたことが判る。これは$\boldsymbol{C}^{(3)}$が系統関係上、後に続くクラスターのない“絶滅”したクラスターであることと強く関連しているものと推測できる。

(5)まとめ

階層ベイズモデルを用いてタンパク質変異の選択圧分布を検出する方法を開発した。空間集積性は事前分布にポッツ模型を採用することで取り込み、熱力学的積分の手法を用いることで超パラメータを全て推定することを可能にした。開発した方法を Flu ウイルスの HA タンパク質に適用することで、選択圧の空間分布を検出した。更に拡張した解析手法を Flu ウイルス（H3N2 型）の 1968 年から 2003 年までの分離株に適用したところ、絶滅したクラスターにおいて、多様化圧がかかっていたアミノ酸残基における置換のほとんどで物理化学的な性質を変えていないことが判明した。

<引用文献>

① S. Yokoyama, et al., Proc. Nat. Acad. Sci. U.S.A., 105, 2008, 13480–13485.

② T. Watabe, et al., Mol. Biol. Evol., 27, 2010, 1782-1791.

③ Z. Yang, et al., Genetics, 155, 2000, 431-449.

④ Y. Suzuki, Mol. Biol. Evol., 21, 2004, 352-2359.

⑤ D.J. Smith, et al., Science, 305, 2004, 371-376.

⑥ B.F. Koel, et ai., Sience, 342, 2013, 976-979.

５．主な発表論文等

〔雑誌論文〕（計３件）

① Teruaki Watabe, Yoshiyasu Okuhara and Yusuke Sagara, A hierarchical Bayesian framework to infer the progression level to diabetes based on deficient clinical data, Computers in Biology and Medicine, **査読有り**, 50, 2014, 107-115

② Teruaki Watabe and Hirohisa Kishino, Spatial distribution of selection pressure on a protein based on the hierarchical Bayesian model, Molecular Biology and Evolution, **査読有り**, 30, 2013, 2714-2722

③ Teruaki Watabe and Hirohisa Kishino, Spatial distribution of selection pressure on a virus protein deriving its adaptation to the environment, Proceedings of the Institute of Statistical Mathematics, **査読有り**, 60, 2012, 305-316

〔学会発表〕（計４件）

①渡部輝明、タンパク質立体構造情報を考慮した突然変異選択圧分布の解析、日本遺伝学会、2014 年 9 月 19 日、長浜バイオ大学（滋賀県）

②渡部輝明、インフルエンザヘマグルチニンタンパク質上における選択圧の分布と変遷の検出、日本進化学会第 16 回大阪大会、2014 年 8 月 21 日、高槻現代劇場（大阪府）

③Teruaki Watabe, Spatial distribution of selection pressure on a protein based on the hierarchical Bayesian model, Annual Meeting of the Society for Molecular Biology and Evolution (SMBE 2013), 2013 **年 7 月 7 日-11 日**, (Chicago, IL, U.S.A.)

④渡部輝明、階層ベイズモデルによるタンパク質にかかる多様化圧の時空間集積性の推定、日本進化学会第 14 回東京大会、2012 年 8 月 21 日、首都大学東京（東京都）

６．研究組織

(1)研究代表者

渡部　輝明（WATABE,Teruaki）
高知大学・教育研究部医療学系・講師
研究者番号：90325415

(2)研究分担者

岸野　洋久（KISHINO,Hirohisa）
東京大学・農学生命科学研究科・教授
研究者番号：00141987