Vol.2009-BIO-18 No.5
2009/9/17

オリゴヌクレオチド組成を用いた一括学習型自己組織化地図法による A 型インフルエンザウイルスの俯瞰的な特徴解明

岩崎裕貴† 池村淑道† 伊藤正恵† 阿部貴志†

A 型インフルエンザウイルスの全ゲノム配列を対象に，オリゴヌクレオチド組成を基にゲノム配列の俯瞰的な特徴把握が可能な BLSOM 解析を行った．宿主・セグメント・サブタイプを反映して分離(自己組織化)しており，これらの分離に寄与している特徴的なオリゴヌクレオチド頻度を知ることができ，宿主や亜型の特徴，感染性や毒性などに関連した特徴，時系列的な変化を効率的に把握することが可能となった．

A novel informatics and visualization method to analyze all influenza A virus genomes on the basis of Batch-Learning Self-Organizing Map

Yuki Iwasaki† Toshimichi Ikemura† Masae Itoh†
Takashi Abe†

We used Batch-Learning Self-Organizing Map (BLSOM) for oligonucleotide frequencies (e.g., tri-, tetra-, and pentanucleotides) to clarify features of influenza A virus genomes. Eight fragments were separated (self-organized) from each other. When we focused on each segment, separation according to host (human, avian, etc.), subtype (H1N1, H2N2, etc.) and epidemic year was observed, visualizing diagnostic oligonucleotides responsible for the separation.

1. はじめに

塩基配列解読技術の急速な進展により，2,000 種を超える生物種のゲノム解読が完了，ないしは進行している．ゲノムサイズの小さいウイルスについては，40,000 を超える株(strain)の配列が既にデータベースに登録された．ウイルスは様々な環境に適応するために，頻繁に変異を起こしており，医学的な重要さから，その変異の多様性や時間的経過を知るために，多数の株(strain)の配列解読が進められている．一般の生物種に比べて，ウイルスは飛躍的に速い進化速度を持ち，刻々とゲノム配列を変化させている．ウイルスゲノムの時系列的な変化を把握することは，一般の生物種では数十万年や数百万万年で起きた分子進化過程を，年単位で，かつ現在進行形で把握することにつながる．一般の生物種の分子進化の研究で得られている知見や解析手法を利用することで，ウイルスゲノムの変化について未来予測を行うとすれば，その検証が近未来に可能となる．

現在，流行している新型インフルエンザウイルスをはじめ，過去に数度のパンデミックを引き起こし，今もなお人類に甚大な被害をもたらしている A 型インフルエンザウイルス(以下，インフルエンザ A ウイルスとする)は，ウイルス種の中でも現時点で特に注目を集めている[1][2]．インフルエンザ A ウイルスは８本の文節ゲノム(セグメント)を持つ RNA ウイルスであり，各セグメントには 1 ないしは 2 個の遺伝子が存在する．亜型間や他の宿主のインフルエンザ A ウイルスとゲノムセグメントの置換を行うことにより，今まで感染できなかった宿主にも感染可能な新型株を生み出す．現在流行している新型インフルエンザ A ウイルスにおいても，これがパンデミックの原因となっており，過去に何度か経験したような甚大な被害を再びもたらす恐れが指摘されている[3]．これを防ぐ，あるいは被害を最小限に抑えるには，インフルエンザ A ウイルスのゲノムの特徴や，その変化について網羅的，かつ俯瞰的に理解をした上で，可能であれば予測をすることが重要になる．

我々は，ゲノムに潜む生物種固有の特徴を解明する目的で，大量かつ多次元データの 2 次元や 3 次元でのクラスタリングと視覚化法として，ヘルシンキ大学のコホネンらが開発した，教師なし学習アルゴリズム「自己組織化マップ法(Self-Organizing Map, SOM)」に着目した[4][5][6]．コホネン SOM の長所を生かしながら，再現性のある分類結果を得るアルゴリズムとして「一括学習型自己組織化マップ法(Batch-Learning Self-Organizing Map, BLSOM)」を開発し，ゲノム配列解析に適用してきた[7][8]．3 連や 4 連塩基といった連続塩基(オリゴヌクレオチド)の出現頻度に着目することで，生物種に関する情報を計算機に与えなくても，断片ゲノム配列の大半が生物種別に高精

† 長浜バイオ大学
Nagahama Institute of Bio-Science and Technology

度に分類（自己組織化）し，ゲノムの特徴を網羅的，かつ俯瞰的に理解することを可能にした[8][9][10][11]．本稿では，これまで開発を行ってきた BLSOM を用いて，新型の登場で猛威をふるっているインフルエンザ A ウイルスゲノムを主対象に解析を行った．セグメントの特徴を反映した高精度な分類が可能であり，さらには宿主（ヒト・トリ・ブタ等）や亜型（H1N1・H2N2・H3N2 等），年代ごとでの分離が見られ，ゲノム配列に内在している多様な特徴を抽出・可視化するがことが可能となった．

2. 方法

2.1 一括学習型の自己組織化マップ法(BLSOM)の概要

コホネンが開発した自己組織化マップ (SOM)は大量で複雑な情報について，似た情報を自ずと集める（自己組織化する）ことを計算機上で実現している[4][5][6]．工学・経済学・言語学のような大量で複雑な情報を解析する分野で普及してきたが，ゲノム塩基配列の解析には殆ど用いられずにきた．コホネンの SOM は出来上がった地図がデータの入力順に依存する問題があった．我々は，従来型の SOM の長所を生かしながら，再現性のある分類結果を得るアルゴリズムに変更するために，「一括学習型の自己組織化マップ法(BLSOM)」を開発してきた[7]．大量データに対する大規模な並列処理が可能となり，大量データの大規模解析に適したアルゴリズムとなった[11]．

連続塩基頻度のみに着目することで，ゲノム断片配列の大半を生物種により高精度に分類（自己組織化）する BLSOM の能力を基に，メタゲノム解析で得られた環境由来の大量ゲノム断片配列について，系統推定法を確立できた [9][10]．この手法は，オルソログ配列セットや配列のアラインメントが不要であり，新規性の高い配列の系統推定が可能である．単一のゲノムに着目すると，ゲノム断片配列は機能部位ごとに分離する傾向を示しており，機能領域の配列に潜む特徴を明らかにできた．BLSOM 法の詳細なアルゴリズムについては，著者らの他の解説書や原著論文を参照されたい[7][8][12][13]. PC クラスタで使用可能な BLSOM を Untrod Inc.と共同で開発している．そのプログラムの使用の希望に関しては，筆者らに連絡を頂きたい．

海水中に豊富に存在するウイルスに着目したメタゲノム解析も報告されている[14]が，ウイルスゲノムには rDNA が存在しないので，系統推定法として広く普及している rRNA 配列に着目した系統推定は不可能である．BLSOM を用いることにより，メタゲノム解析で得られたウイルス由来のゲノム断片配列に対し，オリゴヌクレオチド組成の特徴のみで，ウイルスの系統推定が可能となる．

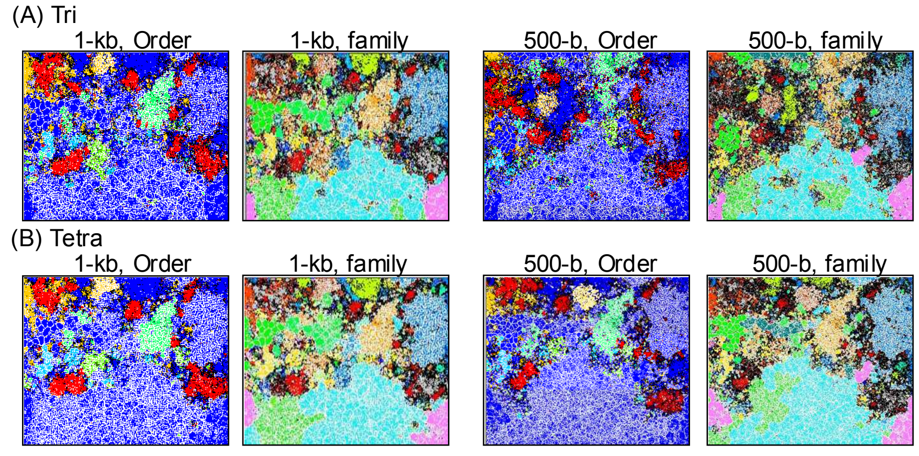

図 1. 全ウイルスを対象にした 3 連続(A), 4 連続(B)塩基での分類結果．ここで，断片化サイズを 1-kb, 500-b とし，Order (目), Family (科)による分類結果を示す．ここで，格子点が同一の生物系統の配列のみからなる場合にはその系統を示す色で，複数の系統の配列が混在する場合には黒色で表す．ここで，色と系統との関係は以下のとおりである．Order: ■; Caudovirales, ■; Herpesvirales, ■; Mononegavirales, ■; Nidovirales, ■; Nidovirales, ■; Orderunclassified. Family: ■; Coronaviridae, ■; Siphoviridae, ■; Hepadnaviridae, ■; Flaviviridae, ■; Poxviridae, ■; Retroviridae, ■; Orthomyxoviridae.

3. 結果と考察

3.1 全ウイルスゲノム配列を対象としたゲノム特徴の俯瞰的把握

広範囲なウイルスゲノムを対象として，BLSOM のクラスタリング能力を検証するために，国際塩基配列データベース上に公開されている全ウイルス種の約 44,000 株の配列を対象に選んだ．各ウイルス株のゲノム配列を，500b と 1kb に断片化した配列を対象に，3 連続ならびに 4 連続塩基頻度について BLSOM 解析を行った(図 1). 図 1 では，ウイルスの order(目)ならびに family(科)と呼ばれる系統レベルでの分離の結果を示した．ここで，格子点が同一の生物系統の配列のみからなる場合にはその系統を示す色で，複数の系統の配列が混在する場合には黒色で表した．大半の格子点が生物系統ごとに色付けされており，生物系統を反映して配列類がクラスタを形成(自己組織化)している．4 連続塩基において，断片化サイズ 1kb と 500b での family レベルでの分離は，

Vol.2009-BIO-18 No.5
2009/9/17

99%と 98%以上と非常に高精度に分類されていた．ウイルス種に固有なゲノム配列の特徴は，500 塩基レベルの短い断片配列においても保持されていることが判明しており，メタゲノム解析で得られたウイルス由来のゲノム断片配列に対しても，その系統推定に有用であることが示された．

3.2 全インフルエンザ A ウイルスのオリゴヌクレオチド組成によるゲノム特徴の俯瞰的把握

現在，流行している新型インフルエンザウイルスをはじめ，過去に数度のパンデミックを引き起こし，今もなお人類に甚大な被害をもたらしているインフルエンザ A ウイルスに着目した．インフルエンザ A ウイルスは，8 つの分節ゲノム (セグメント) を持つが，各セグメント番号とそこに乗っている遺伝子との対応関係は次のとおりである．セグメント 1 は PB2，セグメント 2 は PB1，セグメント 3 は PA，セグメント 4 は HA，セグメント 5 は NP，セグメント 6 は NA，セグメント 7 は M1 と M2，セグメント 8 は NS1 と NS2 となる．最も短いセグメント 8 でも，800 塩基以上あるので，図 1 の BLSOM の結果から，充分な分離能が期待できる．

現在公開されているインフルエンザ A ウイルスの 5350 株 (42,800 ゲノムセグメント) を対象に，3 連続，4 連続，ならびに 5 連続塩基の頻度を対象とした BLSOM 解析を行った．ここでは，インフルエンザ A ウイルスに対し，同一の BLSOM を用いて，セグメントごと(図 2A)と宿主ごと(図 2B)での分類結果を示す．セグメントごとでは，得られた BLSOM マップ上での黒の割合がほとんどなく，非常に高精度に分類されていた．宿主ごとに見た場合，興味深いことに，各セグメントの領域内において，宿主を反映した分離が見られた．例えば，4 連続塩基 BLSOM での場合では，セグメント 2，4 ならびに 6 では，マップの右上付近に隣り合って分類されているが，宿主別での分布を見ると，各セグメントともヒト由来のゲノムが隣り合って分類されており，ヒト特有の領域が形成されている．これは他のセグメントについても同様であり，宿主の違いを反映したオリゴヌクレオチド組成の特徴が存在していた．

また，ヒト以外の宿主の分布を見てみると，トリ由来については，ヒト由来の領域と完全に分離しているのに対し，ブタ由来については，ヒトとトリ由来の領域の境界付近に分布している場合が多かった．ブタ由来については，トリとヒト由来の両方の宿主からでも感染することが知られており，両方の宿主の特徴を反映している結果といえる．ウマ由来についても同様のことが言え，インフルエンザ A ウイルスの宿主適応による感染経路解明の点からも興味深い．

3.3 インフルエンザ A の亜型ごとのゲノム特徴の俯瞰的把握

インフルエンザ A ウイルスは，ヘマグルチニン(HA)とノイラミニダーゼ(NA)が、それぞれ 16 種類および 9 種類見いだされており，抗原性の違いとして容易に識別できる．

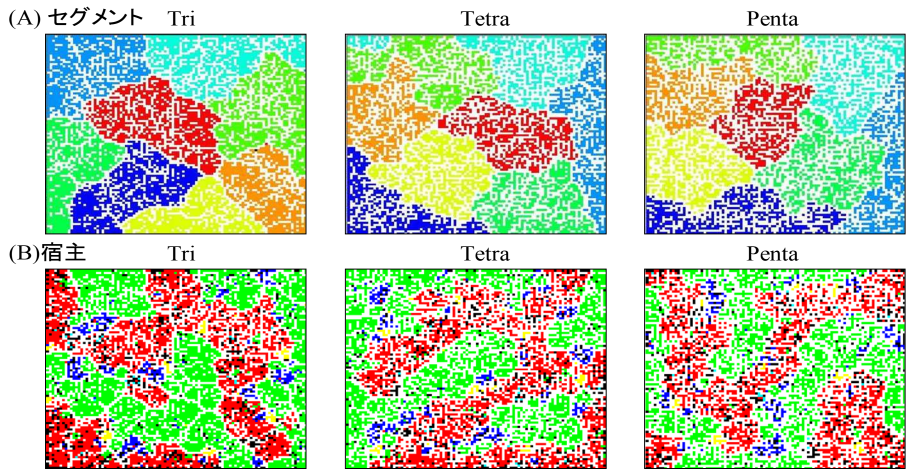

図 2. 全インフルエンザ A ウイルスの全セグメントゲノムを対象にした 3 連続，4 連続，ならびに 5 連続塩基によるセグメント(A)と宿主(B)による分類結果．ここで，色との関係を以下に示す．セグメント: ■，セグメント 1，■，セグメント 2，■，セグメント 3，■，セグメント 4，■，セグメント 5，■，セグメント 6，■，セグメント 7，■，セグメント 8， 宿主：■，ヒト，■，トリ，■，ブタ，■，ウマ，■，その他（アザラシ，トラ etc）．

異なる HA と NA の組み合わせを持つ株が，亜型として分類される．ヒトにおいて，パンデミック（大流行）を引き起こしてきた有名な亜型は，H1N1(スペイン風邪)と H3N2(香港風邪)である．近年，ヒトでの流行が危ぶまれている強毒鳥インフルエンザウイルスの亜型として H5N1 が上げられる．現在，流行している新型インフルエンザ A ウイルスは H1N1 であるが，これまで報告のあった H1N1 とは異なる新型である．

亜型での分離を見るために，ヒトへの感染が知られている H1N1，H3N2，H5N1 に着目した．図 3 に 3 連続，4 連続，ならびに 5 連続塩基でのマップ上の分布を示す．ここで，現在流行している新型 H1N1 インフルエンザウイルス(swH1N1)については，従来からの H1N1 とは別に取り扱っている．従来型の H1N1 と H3N2 については，各セグメントのヒト固有領域について，二分化されており，オリゴヌクレオチド組成の亜型別による特徴の存在が明らかである．一方，現在パンデミックな流行を引き起こしている新型 swH1N1 については，全てのセグメントがこれまでのヒト由来のインフルエンザ A ウイルスとは異なり，トリないしはブタ由来の領域やそれらの境界部位に存在していた．

近年, 流行が危ぶまれている強毒鳥インフルエンザ H5N1 ウイルスについて見ると, 全てのセグメントがトリ由来の領域に分類されていた．これは，現時点ではトリから直接にヒトへ感染した例しかなく，ヒトから分離された株でもトリの特徴のみを有していることを反映している．今後，ヒトでパンデミックを引き起こすかどうかが重要になる．トリの H5N1 株のなかに，現時点でヒトに感染した株よりもさらにヒトの領域に近い部位に位置している株が見られる．各セグメントについても，ヒト由来の領域と近い関係にあるトリの H5N1 株を知ることが重要になる．現時点では，特定の H5N1 株として成立していなくても，セグメント交換で，ヒト由来の特徴により近い強毒株が出現する可能性が考えられる．BLSOM 上で見られるヒトの領域は，ヒト・ヒト感染により大流行を起こした株の特徴を持っている．ヒトの領域に近いことは，ヒト・ヒト感染を起こしやすい性質を持つ可能性を示唆する．今後に注意を払うべきトリの H5N1 株を指摘できると考えている．全てのインフルエンザ A ウイルス株のセグメントを一つのマップ上で俯瞰的に把握し，その特徴を知ることの意義は大きい．

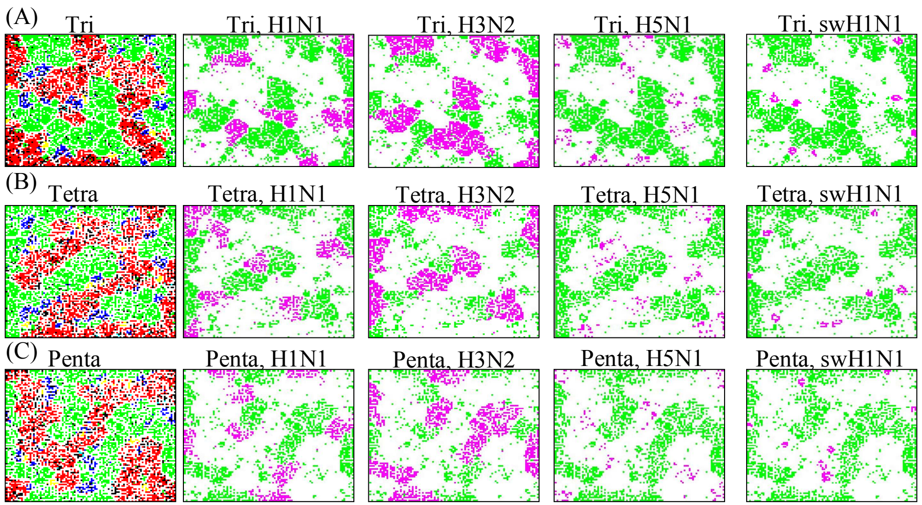

図 3．全インフルエンザ A ウイルスの全セグメントゲノムを対象にした 3 連続 (A)，4 連続 (B)，ならびに 5 連続塩基 (C) による亜型(H1N1, H3N2, H5N1, 新型インフルエンザ A ウイルス(swH1N1))による分類結果．ここで，■によってヒトを宿主とする全データの分類結果と示し，■によって亜型の分類結果を示す．

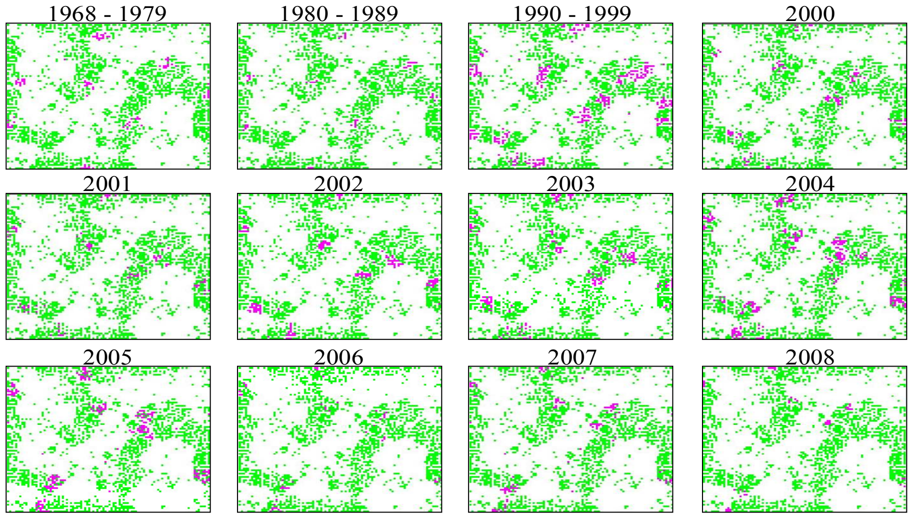

図 4．インフルエンザ A H3N2 ウイルスにおける 5 連続塩基による BLSOM マップ上でのサンプリングされた年代別の分類結果．ここで，■によってヒトを宿主とする全データの分類結果と示し，■によって亜型の分類結果を示す．

3.4 インフルエンザ A ウイルスの変異の年代推移の把握について

インフルエンザ A ウイルスの研究を行う上で，過去にパンデミックを引き起こした亜型の各セグメントの変異の特徴を把握し，今後パンデミックを引き起こす可能性のあるウイルスの特徴を予測することは，この分野の最重要課題の一つである．年代別による変異の特徴を把握するために，図 2 で示した 5 連続 BLSOM マップ上で，ヒトを宿主とする亜型 H3N2 での年代別の分布を見た（図 4）．年代別に，セグメントごとに特徴的な領域が存在しており，年の推移との対応が明らかである．特に，1990 年代と 2000 年については類似性の高い配列を持つ株が流行していたが，2002 年以降では新しい変異を獲得し，現在までに新たな変異を獲得しながら，新たな流行を引き起こしていたことが視覚的に示されている．

図 2 から図 4 に示すような解析を行うことによって，インフルエンザ A ウイルスの各セグメントが，セグメント別，宿主別，亜型別などの様々な特徴を持って，クラスタを形成することが判明した．インフルエンザ A ウイルスの年代別の変化の様子を時系列的に把握できる．さらに，パンデミックを引き起こした株の変遷の様子が視覚的に

把握可能となっており，実験研究者への有用な情報を提供できる強力な解析ツールと考えられる．

3.5 インフルエンザ A ウイルスの宿主による特徴の把握について

これまで，インフルエンザ A ウイルスの全てのセグメントを用いて，BLSOM 解析を行うことにより，各セグメント別の特徴と宿主別の特徴の両方の存在を明らかにすることができた．宿主別の分離は特に興味深く，宿主別の特徴に着目することによって，トリやブタといった異なる宿主からヒトへの感染，それに引き続くヒト・ヒト感染の持続的なサイクルの確立に必要な情報を提供できれば,その意義は極めて大きい．宿主別の特徴をより直接的に把握するために，全てのセグメントのオリゴヌクレオチド頻度を株ごとに足し合わせた後に，3 連続，4 連続，5 連続塩基頻度の BLSOM 解析を行ったところ，宿主別での明瞭な分離が見られた(図 5)．ヒトを宿主とする H1N1, H3N2 において，各々の亜型の領域が完全に分離していたことから，亜型別における塩基組成の違いも識別している．また，ヒトから分離された強毒鳥インフルエンザ H5N1 ウイルスについてみてみると，全てのセグメントがトリ由来の領域に分類されていた．これは，現時点ではトリから直接ヒトにかかった例しかなく，トリ由来の特徴のみを有していることを示している．新型 swH1N1 については，過去に流行していた H1N1 とは異なる特徴を持ち，ブタやトリの境界付近に位置し，ヒトの主領域とは明瞭に分類していた．新型 swH1N1 について，BLSOM を用いて，現時点での特徴とヒトの固有の主領域の特徴とを比較することにより，ヒトでさらに感染力を増す可能性のある変異を予測する上での情報の提供が可能になる．さらには，抗原提示部位での変異を予測すれば，ワクチンのデザインにおいても有用な情報を提供できる．

4. おわりに

BLSOM は教師なしのアルゴリズムであるが，ウイルスゲノムの大半が連続塩基の出現頻度の類似度のみで，予備知識なしに生物種別に分離した．ゲノム配列に潜む各ウイルス種の特徴を明らかにできる．従来の相同性解析や進化系統解析と異なり，オーソログ配列や配列間のアラインメントが必要でなく，連続塩基の出現頻度のみで系統関係を知ることが可能である．従って，rDNA のような生物系統に普遍的なオーソログ配列を持たないウイルスゲノムについては，異なるウイルス間のゲノム比較やその特徴の俯瞰的把握を行うための強力な解析ツールとなる．メタゲノム解析で得られる，ウイルス由来配列の系統推定も可能となる．

インフルエンザ A ウイルスにおいては，セグメントや宿主別によるゲノム配列の特徴の抽出が可能となった．宿主別の特徴が把握できたことは，新しい宿主に対する感染能力の獲得機構の解析やゲノム変異パターンの予言をする上での新規手法を提供することになると考えている．

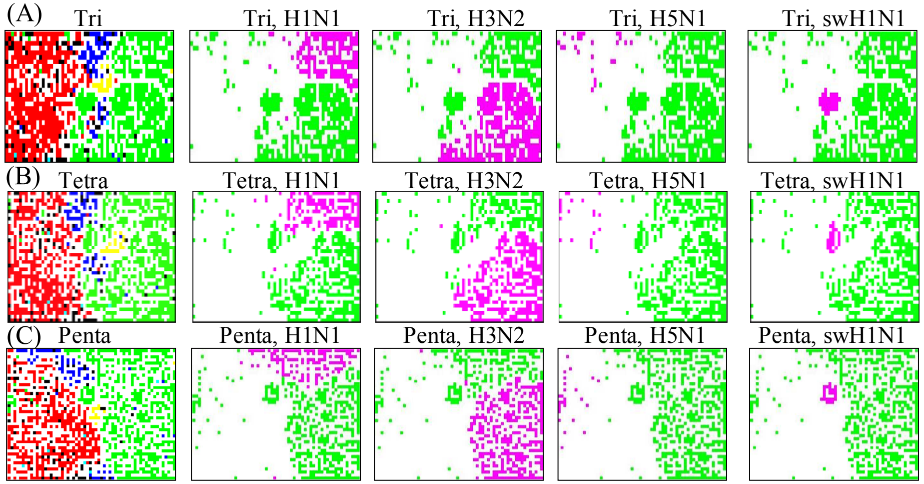

図 5, インフルエンザ A ウイルス株情報を対象にした 3 連続 (A)，4 連続 (B)，ならびに 5 連続塩基 (C) による宿主別による分類結果．ここで，色と宿主情報との関係を以下に示す． ■ ヒト，■，トリ，■，ブタ，■，ウマ，■，その他（アザラシ，トラ）．

謝辞 本報告に関わる調査は，科学研究費補助金（ 基盤研究（C），ならびに，若手（B)）の助成を得て実施したものである．記して厚く御礼を申し上げる．

参考文献

1) 板村繁之: SARS，新型インフルエンザ，トリインフルエンザ : missing link を探して，蛋白質核酸酵素, Vol.49, pp.773-780 (2004).
2) 村本裕紀子，野田岳志，河岡義裕: インフルエンザウイルスの選択的ゲノムパッケージング機構，蛋白質核酸酵素, Vol.51, pp.1596-1601 (2006).
3) Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team.: Emergence of a Novel Swine-Origin Influenza A (H1N1) Virus in Humans, N. Engl. J. Med, Vol.360, pp.2605-2615 (2009).
4) Kohonen, T.: Self-organized formation of topologically correct feature maps, Biol. Cybern., Vol.43,

pp.59-69 (1982).
5) Kohonen, T.: The self-organizing map, Proceedings of the IEEE, Vol.78 pp.1464-1480 (1990).
6) Kohonen, T., Oja, E., Simula, O., Visa, A. and Kangas, J.: Engineering applications of the self-organizing map, Proceedings of the IEEE, Vol.84, pp.1358 – 1384 (1996).
7) Kanaya, S. Kinouchi, M. Abe, T. Kudo, Y. Yamada, Y. Nishi, T. Mori, H. and Ikemura, T.: Analysis of codon usage diversity of bacterial genes with a self-organizing map: characterization of horizontally transferred genes with emphasis on E. coli O157 genome, Gene, Vol.276, pp.89-99 (2001).
8) Abe, T. Kanaya, S. Kinouchi, M. Ichiba, Y. Kozuki, T and Ikemura, T.: Informatics for unveiling hidden genome signatures, Genome Research, Vol.13, pp.693-702 (2003).
9) Abe, T. Sugawara, H. Kanaya, S. Kinouchi, M. Matsuura, Y. Tokutaka, H. and Ikemura, T.: A large-scale Self-Organizing Map (SOM) constructed with the Earth Simulator unveils sequence characteristics of a wide range of eukaryotic genomes, Proceedings of Workshop 2005 on Self-Organizing Maps, pp.187-194 (2005).
10) Uchiyama, T. Abe, T. Ikemura, T. and Watanabe, K.: Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes, Nature Biotechnology, Vol.1, pp.88-93 (2005).
11) Abe, T. Sugawara, H. Kanaya, S. and Ikemura, T.: Sequences from almost all prokaryotic, eukaryotic, and viral genomes available could be classified according to genomes on a large-scale Self-Organizing Map constructed with the Earth Simulator, Journal of the earth simulator, Vol.6, pp.17-23 (2006).
12) 阿部貴志， 上原啓史， 金谷重彦， 池村淑道: 環境由来大量 DNA 配列を利用した難培養性生物群の系統推定のための新規情報学手法， マリンメタゲノムの有効利用(松永是，竹山春子監修)， シーエムシー出版，pp.228-239 (2009).
13) 阿部貴志， 池村淑道， 木ノ内誠， 中村由紀子， 前野聖， 金谷重彦: 自己組織化マップ法・その発展－－医学・生物学からバイオ産業への応用まで， 自己組織化マップ法・その発展－－生物学から社会学まで （徳高平蔵 監修)， シュプリンガージャパン出版， pp.87-98, (2007）.
14) Edwards R.A. and Rohwer F. Viral metagenomics, Nature Review Microbiology, Vol.3, pp.504-510 (2005).