3B8-3

アイテム集合間の相関変化検出による インフルエンザウイルス遺伝子データの解析

An Analysis of Influenza Virus by a method based on differences in correlations between itemsets

谷口剛[*1] 伊藤公人[*2] 五十嵐学[*2] 村上悌治[*2] 原口誠[*1]
Tsuyoshi TANIGUCHI Kimihito ITO Manabu IGARASHI Teiji MURAKAMI Makoto HARAGUCHI

[*1]北海道大学大学院情報科学研究科コンピュータサイエンス専攻
Division of Computer Science, Hokkaido University

[*2]北海道大学人獣共通感染症リサーチセンター
Research Center for Zoonosis Control, Hokkaido University

We anaylize amino acid sequences of influenza viruses by a data mining method based on differences in correlations between itemsets. In order to obtain important knowledge for antigenic drift, we take advantage of domain knowlege of influenza virus. In this paper, we discuss the possibility that interesting knowledge can be found by our method, based on our experiments.

1. はじめに

本研究では，インフルエンザウイルス遺伝子データに対してデータマイニング手法を適用することにより，インフルエンザウイルスの抗原変異における重要な知識を抽出することを目指す．データマイニング手法の適用により，以下のようなことが期待される．まずは，インフルエンザウイルスの過去の抗原変異から何らかの規則が発見できれば，その規則が将来の抗原変異を予測するための手がかりとなる可能性がある．さらに，現状では研究者の経験に基づきインフルエンザウイルスの抗原変異に対する仮説を立て，生物学的な実験を行っている [2] が，もしインフルエンザウイルスの研究者が思いもよらないような規則が発見されれば，その規則から仮説を立て，その仮説を確認するための生物学的な実験を行うことが可能となる．

ここで抗原変異の過程において，もしインフルエンザウイルスが単純な挙動を示すならば，過去の抗原変異からも単純な規則が発見されるだろう．その場合，現状では流行しているウイルスを不活化してワクチンを製造しているのに対して，その発見された規則に基づき，インフルエンザが流行する前に先回りしてワクチンを生成することも可能である．しかし，実際にはインフルエンザウイルスの残基位置間の複雑な相互作用によって，そのような単純明快な規則が発見されることは少ない．通常の知識発見アルゴリズムを適用した場合には，ある残基位置のアミノ酸が *A*(アラニン) であるときに別の残基位置のアミノ酸は *D*(アスパラギン酸) である，というような規則のみが発見されるであろう．インフルエンザウイルスの多くの残基位置はウイルスとしての機能を保持するためにアミノ酸が置換しないため，多くのこのような規則は、将来の変異を予測するためには、ほとんど意味を成さない．

したがって，本研究ではインフルエンザウイルスの領域知識を利用し，単純な手法では埋もれてしまうような局所で成り立つ規則を積み重ねることによって，全体の挙動を明らかにしていくことを目指す．インフルエンザウイルスのアミノ酸残基位置には抗原変異に重要なアミノ酸置換をする残基位置と，ほとんどアミノ酸置換しない残基位置があることがわかっている．そこで例えば，ある残基位置のアミノ酸の置換に注目したときの他の残基位置間のアミノ酸置換の関係の差異を基に浮かび上がってくる関係，もしくは，時間の変化による残基位置間の関係の変化を基に特徴的な関係として顕在化していない潜在的に重要な関係に注目したい．

上記のような観点から，本研究では，インフルエンザウイルスの２つの領域知識を利用する．１つ目は，インフルエンザウイルスは進化の過程において，図 2 にも示されるように，一系統のみ生き残っていくような挙動を示すことである．一般のタンパク質はこのような挙動を示さない．この性質を利用すると，抗原変異において後のウイルス株に引き継がれるようなアミノ酸置換を進化系統樹を利用することにより検出することができる．２つ目は，抗体に認識される部位と考えられている５つの抗原性決定基である．抗原性決定基を考慮し，領域間の隠れた関係を発見することができれば，インフルエンザウイルスの抗原変異のメカニズムを解き明かす上で貢献できる可能性がある．

インフルエンザウイルスの領域知識を利用し，本研究では，以下の３つの観点で計算機実験を行った．1) ある領域の残基位置のアミノ酸が置換するときの他の領域の残基位置のアミノ酸の置換の有無，2) 抗原変異の流れの中での 1) の関係の変化の有無，3) ある領域の残基位置の置換に注目したときの１の関係の差異の有無，である．1) では，異なる領域の残基位置のアミノ酸が共に置換することがあるかどうかを明らかにする．2) では，1) の関係が時期によって変化するかどうかを明らかにする．3) では，ある領域のアミノ酸置換が，他の領域間のアミノ酸置換に影響を及ぼすことがあるかどうかを明らかにする．

2. インフルエンザウイルス遺伝子データ解析の準備

2.1 インフルエンザウイルス

毎年世界中でインフルエンザが流行し，発熱，急性肺炎等の重篤な疾病を引き起こしている．インフルエンザウイルスの遺伝子は突然変異を起こしやすく，毎年ごく僅かな変異ウイルスが人の免疫システムから逃れて生き残り，その翌年，抗原性が少し異なる変異ウイルスとして流行を繰り返す．一般に，ウイルス感染症の予防には，薬品によって不活化したウイルスをワ

連絡先: 谷口剛，北海道大学大学院情報科学研究科，〒 060-0814 札幌市北区北 14 条西 9 丁目，TEL(FAX):011-706-7161, E-mail:tsuyoshi@kb.ist.hokudai.ac.jp

クチンとして注射する手法が有効である．しかし，インフルエンザの場合ウイルスの抗原性が変化し続けるため，ワクチンによる効果が長続きしない．タミフル等の抗ウイルス剤も使用されるが，薬剤耐性ウイルスへの変異も報告されている．これらの理由から，決定的な予防法の開発は難しいとされる．

インフルエンザウイルスは，抗原変異によって流行を繰り返す．一般に，ウイルスに感染した個体の体内では，ウイルスの抗原領域を特異的に認識する抗体が産生される．抗体はウイルス表面のタンパク質に結合することで，ウイルスが細胞に侵入する機能を阻害して，細胞でウイルスが複製されることを抑制する．一方，集団においてウイルスは咳やクシャミによって別の個体に感染する．この感染を繰り返すうちに，突然変異によって，抗体が認識する部位のアミノ酸が偶然別のアミノ酸に置き換わったウイルスが現れる．この変異ウイルスが以前の感染時に産生された抗体によって病原体として認識されなければ，一度インフルエンザに罹ったことがある個体にもまた感染する．このように、ウイルスが抗体との結合から逃れていくためのアミノ酸の変化を抗原変異と呼ぶ．

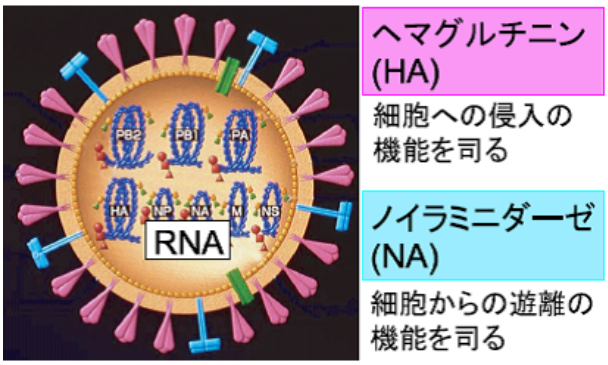

図 1: インフルエンザウイルスの構造模式図

インフルエンザウイルスは赤血凝集素 (HA) とノイラミニダーゼ (NA) の二種類のスパイクタンパク質を粒子表面に持つ (図 1)．HA および NA はウイルスの表面に突出しているため，抗体による免疫システムの標的となりやすい．インフルエンザの抗原変異は，HA および NA 上のアミノ酸が突然変異によって別のアミノ酸に置換され，タンパク質の部分的構造が変化し，抗体によって認識されなくなることに起因する．HA と NA はそれぞれ宿主細胞への侵入と細胞からの遊離を司る。HA と NA はこれらの機能を保持する必要があり，抗原変異におけるアミノ酸置換には何らかの制限があると考えられる．そのため，インフルエンザウイルスの抗原変異におけるアミノ酸置換に，ある種の規則性が潜在する可能性がある．

2.2 データ

本研究では，香港風邪の病原体ウイルスである H3N2 亜型インフルエンザウイルスの HA タンパク質における解析をするため，HA タンパク質のアミノ酸配列を，NCBI Influenza Virus Resource[3] からダウンロードした．そして，多重配列アラインメントによって，長さの異なるアミノ酸配列を整列した．H3N2 亜型ウイルスの HA タンパクのアミノ酸配列の長さは約 550 であり，約 330 の長さの HA1，約 220 の長さの HA2 の二つの領域を持つ．抗原変異は，1) 主に HA1 領域で起こること，2)HA1 領域のみデータベースに登録されている配列が多数あることから，本実験では，多重配列アラインメントから切り出した HA1 領域のアミノ酸配列 2183 本を解析対象とした．

2.3 進化系統樹を利用したアミノ酸置換の抽出

本研究では，抗原変異におけるアミノ酸置換に注目する．そのために，ウイルス株同士の対応関係を明らかにし，対応するアミノ酸配列の差分をとることによってアミノ酸置換を抽出する．NCBI からダウンロードしたデータにおけるそれぞれのウイルス株には，そのウイルス株が出現した年の情報が含まれており，時系列データとみなすこともできる．結果として多くの場合，インフルエンザウイルスの進化と時系列情報が対応するが，基本的にはそれぞれの年において様々な残基位置のアミノ酸が置換したウイルス株が混在し，ウイルス株が生き残るようなアミノ酸置換をしなかったため，次の年のウイルス株と進化における対応関係のないウイルス株も存在する．また，アミノ酸配列が同じウイルス株が数年たってから再び出現することもある．したがって，時系列情報のみに基づくウイルス株の対応関係は，インフルエンザウイルスの進化の対応関係として正確ではない．

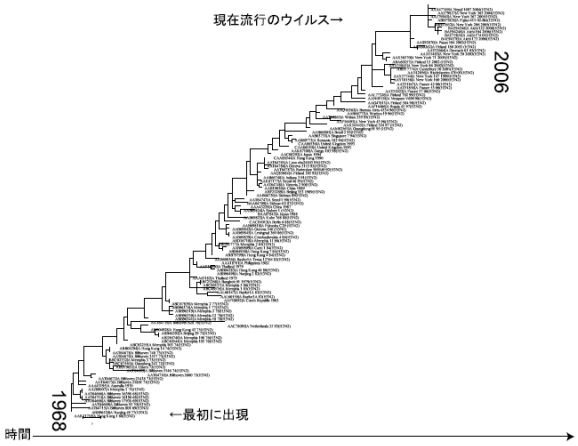

図 2: インフルエンザウイルス (H3N2 亜型) の進化系統樹

本研究では，アミノ酸配列の祖先関係を推定するため，進化系統解析を行った．図 2 に H3N2 亜型インフルエンザウイルスの進化系統樹を示す．一般に，進化系統樹の推定には，近隣結合法が多く用いられる．本研究では，仮想的な祖先 (進化系統樹の葉以外のノード) のアミノ酸配列を推定するために，進化系統解析には最節約法を用いる．進化系統樹の葉以外のノードを内点と呼ぶ．仮想的な祖先は内点となる．

進化系統樹のデータ構造は，図 3 に示すように，木構造である．最節約法を用いて作成される進化系統解析では，内点のアミノ酸配列が推定されるため，木の全てのノードにはアミノ酸配列が対応付けられる．つまり，各ノードはウイルス株および仮想的な祖先ウイルスのアミノ酸配列に対応する．根は，新型インフルエンザとして出現した株である．幹はインフルエンザウイルスの過去の抗原変異の履歴を表す．幹の末端は，現在流行しているウイルス株である．枝は，突然変異によるアミノ酸置換に対応する．各ノードにはアミノ酸配列が対応付けられているので，枝には具体的なアミノ酸置換の集合を対応付けることが可能である．図 3 中において，2 番目のアミノ酸 D が G に置換したことを $D2G$ と表記している．

ここで，**アイテム**を定義する．アミノ酸の集合 A を $\{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, Y\}$ で表す．進化系統樹上のある枝 e に割り当てられたアミノ酸置換の集合 T_e を $T_e = \{x_1n_1y_1, x_2n_2y_2, \cdots, x_kn_ky_k\}$

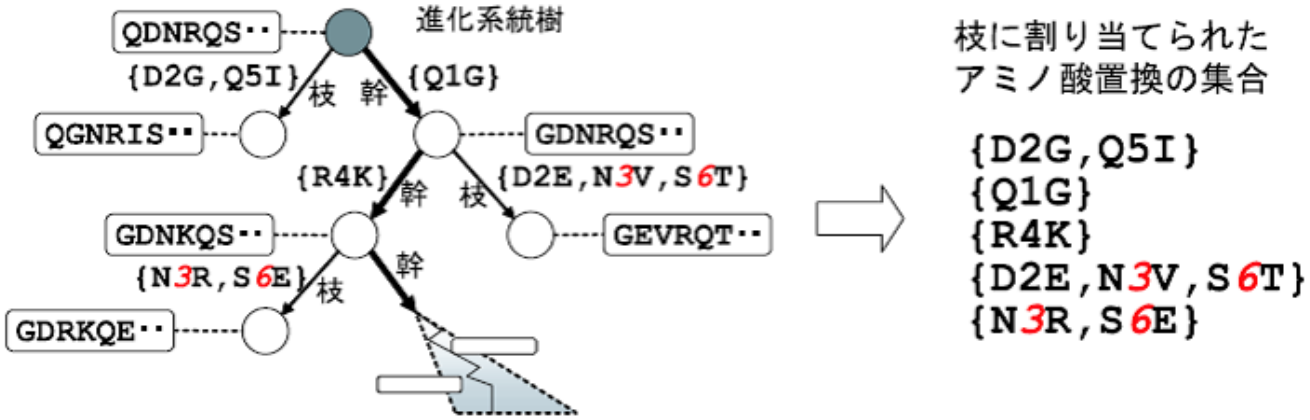

図 3: 進化系統樹からのアミノ酸置換の抽出

とする．ここで，$x_1, \cdots, x_k, y_1, \cdots, y_k$ はアミノ酸であり，$n_1, \cdots, n_k$ は，アミノ酸残基位置である．このとき，枝 e においてアミノ酸置換が起こったアミノ酸残基位置の集合 S_e を $S_e = \{n_1, n_2, \cdots, n_k\}$ とし，本稿の実験におけるアイテムとする．つまり，本稿の実験においては，残基位置間の関係に注目して解析を行う．

2.4 アミノ酸残基の共変異

一般のタンパク質の進化過程において．二つ以上の異なる位置のアミノ酸が共に変異する場合が知られている．二つ以上のアミノ酸が共に置換する進化過程を，**アミノ酸残基の共変異**と呼ぶ．アミノ酸残基の共変異は，タンパク質の構造や機能の保持に重要な突然変異であると考えられている．一つの自然な解釈として，ある位置でのアミノ酸変異がタンパク質の構造上の歪みを引き起こし，別の位置のアミノ酸変異がこの歪みを修正する現象であることが挙げられる．

本稿においては，アミノ酸残基の共変異は，枝 e においてアミノ酸置換が起こったアミノ酸残基位置の集合 S_e の部分集合 $S \subseteq S_e$ に対応する．S のことを**アイテム集合**と呼ぶ．

本研究では，既に頻出アイテム集合導出に基づく共変異の検出を試みており，いくつかの興味深い結果を得ている [1].

2.5 抗原性決定基

インフルエンザウイルスの HA 分子上には，図 4 に示すような 5 つの抗原領域（A, B, C, D, E）が存在すると考えられている．図 4 において，ピンクが A，青が B，緑が C，黄色が D，茶色が E に対応する．

抗原変異に関連のあるアミノ酸置換はほぼ上記の領域で生じる．ここで明示的には現れていない領域間の隠れた関係を見つけることができれば，インフルエンザウイルスの抗原変異における隠れたメカニズムを解き明かす情報となる可能性がある．本研究では，抗原性決定領域を考慮し，共変異の差異に基づき領域間の隠れた関係を検出することを試みる．

以下にそれぞれの領域の残基位置番号を示す．

A 122, 124, 126, 127, 130, 131, 132, 133, 135, 137, 138, 140, 142, 143, 144, 145, 146, 150, 152, 168.

B 128, 129, 155, 156, 157, 158, 159, 160, 163, 164, 165, 186, 187, 188, 189, 190, 192, 193, 194, 196, 197, 198.

C 44, 45, 46, 47, 48, 50, 51, 53, 54, 273, 275, 276, 278, 279, 280, 294, 297, 299, 300, 304, 305, 307, 308, 309, 310, 311, 312.

D 96, 102, 103, 117, 121, 167, 170, 171, 172, 173, 174, 175, 176, 177, 179, 182, 201, 203, 207, 208, 209, 212, 213, 214, 215, 216, 217, 218, 219, 226, 227, 228, 229, 230, 238, 240, 242, 244, 246, 247, 248.

E 57, 59, 62, 63, 67, 75, 78, 80, 81, 82, 83, 86, 87, 88, 91, 92, 94, 109, 260, 261, 262, 265.

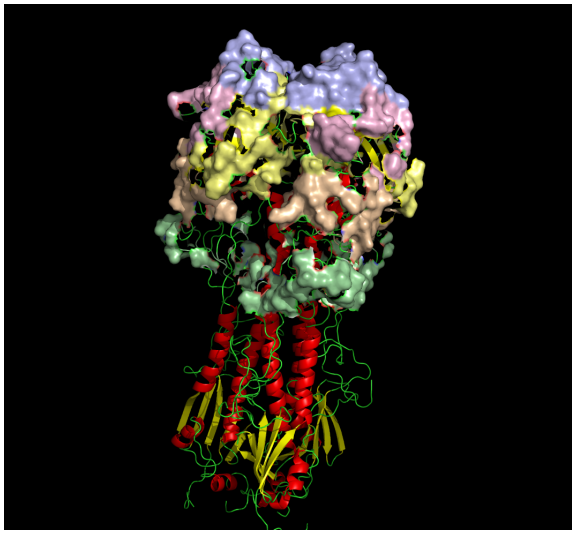

図 4: 抗原性決定基

3. 抗原性決定基を考慮した共変異の変化・差異の検出

3.1 抗原性決定基間の共変異

まずは，抗原性決定基をまたいだ共変異が存在するかどうかを確かめるための実験を行った．本実験では，進化系統樹を

利用して検出したアミノ酸置換の集合を基に，それぞれの株のどの領域が置換したかを識別した．そして，全体のウイルス株に対するある領域 X のいずれかの残基位置のアミノ酸が置換したウイルス株の割合を $P(X)$，X と領域 Y のいずれかの残基位置のアミノ酸が，X と Y で同時に置換したウイルス株の割合を $P(XY)$ と表記する．そのときに，X と Y の共起の度合いを $P(XY)/P(X)P(Y)$ で評価した．この式は log をとらない自己相互情報量である．

実験結果

A, B, C, D, E のほとんど全ての 2 つの組み合わせが，高い共起の度合いを示した．特に領域 C と領域 E の共起の度合いは他の領域間の共起と比べて高かった (P(CE)/P(C)P(E)=1.60)．以上の結果より，抗原性決定基をまたいだ共変異は存在することを確認した．

3.2 抗原性決定基間の共変異の時系列による変化

前副節の結果により，抗原性決定基間の共変異は存在することを確認したが，ほとんど全ての組み合わせが高い共起を示したので，注目すべき共変異は発見できなかったという言い方もできる．領域をまたいで共変異することはよく起こるということを確認したに過ぎない．そこで，時系列の流れの中で抗原性決定基間の共変異が変化するのか，確認するための実験を行った．本実験ではウインドウの区切り方，その調べ方は様々な方法が存在するが，本実験においては簡単のために，考える期間を 1968 年〜1979 年，1980 年〜1989 年，1990 年〜1999 年，2000 年〜2006 年とし，ウインドウをずらさずにそれぞれの期間における共変異を調べた．

実験結果

特筆すべき結果は，領域 A と C，A と E，C と E の共変異の度合いが 1980 年〜1989 年に急上昇し，その後急下降する劇的な変化である．それ以外は，多少の変化は見られるものの，ほぼ安定した共起が観察される．この結果は，進化の流れに影響を受ける共変異と進化の流れに依存しない共変異が存在する可能性を示しているのかもしれない．このことに関しては，現在調査中である．

3.3 抗原性決定基のアミノ酸置換に注目した共変異の差異

最後に，ある抗原性決定基のアミノ酸置換が，他の領域の共変異に影響を与えるような可能性のある結果が得られるか，実験を行った．つまり，もし影響を与えることがあるならば，少なくともそのアミノ酸置換に注目したときとしないときとでデータ上は共起の度合いに違いが生じるはずである．もちろん劇的な違いが発見されたとしても，それがある領域のアミノ酸置換がほかの領域の共変異に影響を与えることがあることを示すわけではないことに注意する．3.1 の結果と，ある領域 X のいずれかの残基位置のアミノ酸が置換した条件化における，別の領域 Y と Z の共起の度合い $P(YZ|X)/P(Y|X)P(Z|X)$ を比較することによって，実験を行った．$P(Y|X)$ は領域 X のいずれかの残基位置のアミノ酸が置換するウイルス株の中で Y のいずれかの残基位置のアミノ酸が置換する割合を表す．

実験結果

この実験によって，特筆すべき結果を得ることはできなかった．もし，他の領域に影響を与えるような領域が存在するなら，それを考慮したときとしないときで大きな共変異の割合の差異となって現れることを期待したが，ほとんどの組み合わせで大きな差異は発見できなかった．

4. まとめと今後の課題

本研究では，インフルエンザウイルスの遺伝子データの解析において，進化系統樹と抗原性決定基という領域知識を考慮し，共変異の差異 (変化) を調べる実験を行った．抗原性決定基をまたいだ共変異を発見することはできたが，共変異の差異を基に隠れた関係を発見する段階まではいたっていない．現在はそのことを目指して研究を進めている段階である．

参考文献

[1] 谷口 剛，伊藤 公人，五十嵐 学，村上 悌治，高田 礼人，原口 誠，インフルエンザウイルスの進化における共変異の変化の解析，情報処理学会研究報告, 2006-BIO-7, pp. 185 – 192, 2006.

[2] 中島捷久，間断なき流行と新型ウイルス出現の機構，日本臨牀，61 巻 11 号, pp. 1897–1903, 2003.

[3] National Center for Biotechnology Information, ”Influenza virus resource”, Jun 23, 2005. Available from: http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html

[4] R.M. Bush, C.A. Bender, K. Subbarao, N.J. Cox, and W.M. Fitch. Predicting the Evolution of Human Influenza A. Science, 286(5446): 1921–1925, 1999.

[5] D.C. Wiley, I.A. Wilson, JJ Skehel, et al. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature, 289(5796):373–378, 1981.

[6] D.J. Smith, A.S. Lapedes, J.C. de Jong, T.M. Bestebroer, G.F. Rimmelzwaan, A.D.M.E. Osterhaus, and R.A.M. Fouchier. Mapping the Antigenic and Genetic Evolution of Influenza Virus. Science, 305(5682):371–376, 2004.

[7] J.B.Plotkin, J. Dushoff, and S.A. Levin. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. PNAS, 99(9):6263–6268, 2002.