社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－67 (56)
2007－BIO－11 (56)
2007／12／21

グループ列分割に基づくインフルエンザウイルスのアミノ酸置換における時代的変化の解析.

谷口剛[1] 原口誠[1] 伊藤公人[2]

北海道大学，大学院情報科学研究科[1] 人獣共通感染症リサーチセンター[2]

概 要

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い．抗体による免疫圧力によって，抗原性が変化したウイルスのみが生き残り，次の流行を引き起こす．ウイルスの進化の詳細を理解するために，本論文では，インフルエンザウイルスの進化において，アミノ酸置換の起こる残基位置が時代と共に変化するか否かを明らかにすることを目的とする．コントラストセットマイニングの枠組みを用いて，隣接するグループ列間の特徴的違いを発見するアルゴリズムを提案し，進化系統解析と提案手法を組み合わせることによって，アミノ酸置換の起こる残基位置の時代的変化を解析する．

A Method for Segmentation of Ordered Groups in a Phylogenic Tree Constructed from Influenza Virus Gene Sequences

Tsuyoshi TANIGUCHI[1], Makoto HARAGUCHI[1] and Kimihito ITO[2]

Graduate School of Information Science and Technology[1],Hokkaido University,
Research Center for Zoonosis Control[2], Hokkaido University

Abstract

The influenza viruses undergo antigenic drift to escape from antibody-mediated immune pressure. In order to predict possible structural changes of their molecules in future, it is important to analyze the patterns of amino acid substitutions in the past. In this paper, we present a method to extract segment of ordered groups in a phylogenic tree constructed from influenza virus gene sequences. We develop an algorithm for segmentation of ordered groups based on a contrast set, which identifies differences between two groups. We apply our algorithm to given ordered groups obtained from the phylogenic tree.

1 はじめに

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 [12, 14]．インフルエンザウイルスの抗原変異は，ウイルスの表面タンパク質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，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．ウイルスの表面タンパク質は主に宿主細胞への侵入を司り，これらの機能を保持する必要がある．このため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，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．本研究では，ウイルスの進化の詳細を理解するために，インフルエンザウイルスの進化において，アミノ酸置換の起こる残基位置が時代と共に変化するか否かを明らかにすることを目的とする．

一方，与えられたいくつかのタプルの集合を比較し特徴を抽出することは，データマイニングの研究領域において重要な課題である．この課題に対し，比較しているデータベースの違いを代表するようなアイテム集合であるコントラストセットを発見する手法が提案されている [2]．コントラストセットとは，あるデータベースにおけるアイテム集合の出現確率が，比較している別のデータベースにおけるアイテム集合の出現確率と大きく異なるアイテム集合である．その出現確率の違いによって，データベースの違いを識別することができる．

コントラストセットの従来研究 [2, 4, 6, 7, 9] において，与えられたデータベースの組に対して，コントラストセットを発見するために，その2つのデータベースを単純に比較する．例えば，データマイニングの研究領域でよく用いられるマッシュルームデータセット [5] にコントラストセットマイニングアルゴリズムを適用するとき，毒性のあるマッシュルームの集合と食用のマッシュルームの集合を比較する．マッシュルームデータセットの場合，毒性の有無によって比較すべきデータベースを簡単に選択することができる．

インフルエンザウイルスの場合は，突然変異によって抗体から逃れたウイルスが次の流行を引き起こし，ヒトの集団内で抗体が産生されたウイルスにおいては流行が終息する．このため，ある期間において危険であったウイルスが，数年後に危険でないウイルスとなり，比較すべきデータベースの組を単純に選択できない．このような場合，全ての可能なタプルの集合の組に対して，コントラストセットを発見するために比較を繰り返さなければならない．そのため，ある集団の特徴を発見するために，適切に比較すべき集団を選択することが必要である．

比較すべき集団の組を発見するために，本研究では，アイテム集合の出現確率の違いによって，与えられたグループ列 $G_1, G_2, \ldots, G_n$ を部分グループ列に分割し，隣接する部分グループ列同士を比較することを考える．本論文では，コントラストセットの出現確率を基に，グループの列を H_k と T_k の2つの部分に分割することができるような全ての3つ組 (X, H_k, T_k) を発見する問題に扱い，効率的な探索を実現できる枝刈りを開発する．

与えられたグループ列をコントラストセットに基づき部分グループ列に分割するためには，コントラストセットを探索するための難しさに加えて，多くの全ての可能なグループ列の分割の候補を扱う難しさに対処しなければならない．コントラストセットの従来研究 [2, 9, 6] では，既にコントラストセットを発見するための効率的なアルゴリズムが提案されている．しかし，本研究における問題では比較すべきグループの集合が動的に変化するため，従来研究の成果によって本研究の問題の効率化をはかることが難しい．つまり，アイテム集合の出現確率の差は，グループ列の分割において非単調に変化する．一方，それぞれのグループにおけるアイテム集合の確率は，集合の包含関係に関して単調に変化する．この単調性に基づき，本研究では，効率的なアルゴリズムを開発した．本研究のアルゴリズムをインフルエンザウイルス遺伝子配列データに適用し，進化系統解析と組み合わせることによって，アミノ酸置換の起こる残基位置の時代的変化を解析する．

2 問題定義

この節では，本研究における**コントラストセットに基づくグループ列分割問題**を定義する．まずはじめに準備としていくつかの定義を与える．

$\mathcal{I} = \{i_1, i_2, \cdots, i_l\}$ を**アイテム**の集合とする．**アイテム集合** X は**サイズ** $|X| = n$ のアイテム $\mathcal{I}$ の部分集合である．**グループ**はそれぞれがユニークなアイテム集合 $t_j \in I (1 \leq j \leq m)$ の集合 $G = \{t_1, ..., t_m\}$ と定義する．それぞれの $t \in G$ は G の**トランザクション**とも呼ばれる．$X \subseteq t$ である場合，t はアイテム集合 X を**含む**という．

$G_1, G_2, \ldots, G_n$ を順序づけられたグループの集合とする．グループの列 $G_i, G_{i+1}, \ldots, G_j$ $(0 \leq i \leq j \leq n)$ **セグメント**と呼ぶ．あるセグメント $G_i, G_{i+1}, \ldots, G_j$ に対し，位置 k $(i < k < j)$ における**ヘッドセグメント** H_k を $G_i, G_{i+1}, \ldots, G_{k-1}$，位置 k における**テイルセグメント** T_k を $G_k, G_{k+1}, \ldots, G_j$ と定義する．

アイテム集合 $X \subseteq \mathcal{I}$ とセグメント S に対し，$\{t \mid t \in G_i, G_i$はS中のグループ，かつ $X \subseteq t\}$ であるようなトランザクションの集合を $O(X,S)$ と表記する．また，S において X を含むトランザクションの出現確率を $P(X,S) = |O(X,S)|/|O(\phi,S)|$ と表記する．

与えられた順序つきグループ集合 $G_1, G_2, \ldots, G_n$ のセグメント S に対し，コントラストセットに基づく順序グループ列分割問題の目的は，アイテム集合 X を基にヘッドセグメント H_k とテイルセグメント T_k を区別できるような (X, H_k, T_k) を発見することである．コントラストセットの定義 [2] に従い，X と H_k，T_k に対し，$diff(X, H_k, T_k)$ を以下のように定義する．

$$diff(X, H_k, T_k) = |P(X, H_k) - P(X, T_k)|.$$

本研究における目的を達成するために，以下のような条件を満たす (X, H_k, T_k) を発見する問題を定義する．

問題定義

与えられた閾値 δ に対し，コントラストセットに基づくグループ列分割問題は，$diff(X, H_k, T_k) \geq \delta$ を満たすような全ての３つ組 (X, H_k, T_k) を発見する

3 アルゴリズム

この節では，コントラストセットに基づくグループ列分割問題のためのアルゴリズムを提案する．本研究の問題において，従来のコントラストセットマイニング問題の成果を利用することは難しい．そこで，それぞれのグループにおけるアイテム集合の出現確率の単調性を利用する．

3.1 素朴な手法

与えられたグループ列 $G_1, G_2, \ldots, G_n$ のセグメント S から条件を満たす３つ組 (X, H_k, T_k) を発見するためにいくつかのアプローチを考えることができる．その中で (X, H_k, T_k) を発見するための１つの素朴な方法として以下のようなものが挙げられる．

1. $|O(X, G_j)|$ と $|G_j|$ を計算する．
2. $|O(X, G_j)|$ と $|G_j|$ に基づき，$diff(X, H_j, T_j)$ を計算する．
3. $diff(X, H_j, T_j) \geq \delta$ を満たす (X, H_j, T_j) を出力する．
4. 全ての可能なアイテム集合とセグメントに対し，上記の過程を繰り返す．

上記の手続きによって，全ての求める３つ組 (X, H_k, T_k) を確実に発見することができる．しかし，可能なアイテム集合とセグメントの数が多い場合，上記の手続きによって (X, H_k, T_k) を得ることは非現実的となる．

また，(X, H_k, T_k) を発見するための別の方法として，コントラストセットマイニングの従来手法の成果を利用できる可能性もある．しかし，本研究の問題においてアイテム集合 X に対して比較すべきグループの集合は変化するので，(X, H_k, T_k) を繰り返し調べなければならない．つまり，コントラストセットマイニングの成果を利用するためには，それぞれのアイテム集合の比較すべきデータベースの候補に対し，何度もコントラストセットマイニングアルゴリズムの適用を繰り返さなければならない．

3.2 アルゴリズムの効率化

この副説では，効率的に (X, H_k, T_k) を発見するための枝刈り規則について説明する．ここでは，特に，セグメントが固定されている場合の枝刈り規則について議論する．アイテム集合 X，ヘッドセグメント H_k とテイルセグメント T_k に対し，$diff(X, H_k, T_k)$ は集合の包含関係に関して非単調に変化する．同様に，それぞれの H_k と $T_k(1 \leq i \leq k \leq j \leq n)$ の組に対して，$diff(X, H_k, T_k)$ は非単調に変化する．一方，それぞれのグループ $G_j(1 \leq j \leq n)$ において，$P(X, G_j)$ は単調に変化する．本研究では，効率的な (X, H_k, T_k) の探索を実現するために，この単調性を利用する．$X \subseteq X'$ とそれぞれのグループ $G_j(1 \leq j \leq n)$ に対し，$P(X', G_j) \leq P(X, G_j)$ が成り立つ．そのときに，H_k と T_k に対し，$O(X, H_k) = \sum_{j=1}^{k-1} O(X, G_j)$ かつ $O(X, T_k) = \sum_{j=k}^{n} O(X, G_j)$ であるので，$P(X', H_k) \leq P(X, H_k)$ と $P(X', T_k) \leq P(X, T_k)$ が成り立つ．(H_{max}, T_{max}) を $diff(X, H_k, T_k)$ の値が最大であるセグメントの組とする．$P(X, T_{max}) \leq P(X, H_{max})$ である場合，以下の関係が成り立つ．

$$0 \leq diff(X', H_k, T_k) \leq P(X, H_{max}).$$

$P(X, H_k) \leq P(X, T_k)$ である場合，以下の関係が成り立つ．

$$0 \leq diff(X', H_k, T_k) \leq P(X, T_{max}).$$

したがって，以下のような枝刈り規則を利用することができる．

枝刈り規則

- $P(X, T_{max}) \leq P(X, H_{max})$，$P(X, H_{max}) < \delta$ であるならば，X' は調べる必要がない．．
- $P(X, H_{max}) \leq P(X, T_{max})$，$P(X, T_{max}) < \delta$ であるならば，X' は調べる必要がない．

上記の枝刈り規則を，最近のデータマイニング研究におけるアイテム集合マイニング問題においてよく用いられることが多い SE-tree(set enumeration tree) というデータ構造を深さ優先的に探索していくアルゴリズム [3, 8] において，あるアイテム集合に対しその上位集合を調べる必要があるかどうかの判定条件として実装し，システムを構築した．

4 実験

4.1 データセット

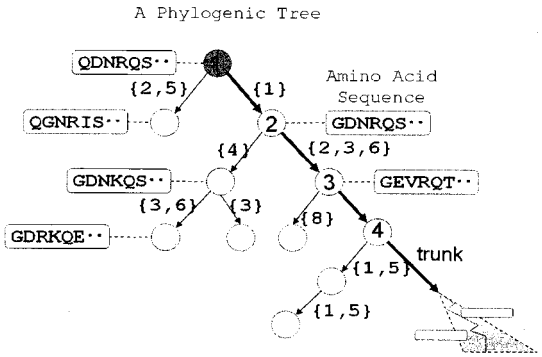
A 型インフルエンザウイルスは，ヘマグルチニンと呼ばれる表面タンパク質のアミノ酸が突然変異によって置換することにより徐々に抗原性が変化する [12, 14, 10]．抗原性の変化によって以前の感染やワクチン接種時に産生された抗体がウイルスに結合することができなくなり，ヒトにおけるインフルエンザの流行の原因となる．

アミノ酸が置換する残基位置が時代と共に変化するか否かを調べるため，前節において述べた手法に基づき，ウイルス遺伝子のデータ分割を行った．

実験において用いたデータセットは 1968 年から 2007 年にヒトから分離された H3N2 の亜型のインフルエンザウイルスの HA 遺伝子の 2737 本の塩基配列からなる．このデータセットは，NCBI Influenza Resources [13] からダウンロードした．それぞれの配列は 984 塩基とそれを翻訳した 328 アミノ酸残基からなる．進化系統解析ソフトウェア Phylip (version. 3.66)[11] の最節約法のルーチンを用いて，塩基配列から進化系統樹を推定した．

推定された進化系統樹は，2737 個の葉と 938 個の内点を含む 3675 個の節によって構成される．A 型インフルエンザウイルスの HA の進化系統樹は，極端に長い幹を一本だけ持つ．。ここで，系統樹の幹は，塩基置換の合計数に関して系統樹の根から末端の葉まで最も長い経路と考える [10]．

938 個の内点のうち，ちょうど 100 個が系統樹の幹上にあった．それぞれの節には，最節約法のルーチンによって，アミノ酸配列がラベルとして付与される．本研究では，辺で連結された 2 つの節のラベルであるアミノ酸配列を比較するという単純な方法によって，アミノ酸置換の集合と系統樹のそれぞれの辺を関連付けた．



group	transaction
G_1	$\{\{2,5\}\}$
G_2	$\{\{4\},\{3,6\},\{3\}\}$
G_3	$\{\{8\}\}$
G_4	$\{\{1,5\},\{1,5\}\}$
⋮	⋮

図 1: 進化系統樹とアイテム集合

系統樹の辺におけるアミノ酸残基位置の集合 $\{\pi_1,\ldots,\pi_n\}$ をトランザクションとする (図 1)．例えば，2 番と 5 番の 2 つのアミノ酸が 1 本の辺において他のアミノ酸に置換している場合，そのトランザクションを $\{2,5\}$ とする．幹上の節 $\{tr_1,\ldots,tr_{100}\}$ のそれぞれの節 tr_i に対して，tr_{i+1} を含まない部分木中の全ての辺に関連付けられたトランザクションの集合を，tr_i のグループとする．例えば，幹上の節　に対応するグループはトランザクションの集合 $\{\{4\},\{3,6\}\{3\}\}$ である．系統樹における全ての幹の節からグループを作成した結果，100 個のトランザクションのグループ $\{G_1,\ldots,G_{100}\}$ が得られた．

4.2　幹上の節とウイルス株の年代の関係

各グループにおけるの関係を図 2 に示す．図 2 において，X 軸はウイルス株の年代，Y 軸は各グループに含まれる各年代の株の数，Z 軸は幹上の節 ID を示す．

図 2 より，系統樹の根に近い (ノード ID が小さい) 節ほど古い年代の株を含み，系統樹の幹の末端に近い節 (ノード ID が大きい) ほど年代の新しい年代の株を含むことがわかる．このことより，本実験におけるグループの順序関係と株の時系列の関係はほぼ対応していることがわかる．したがって，本実験におけるグループの順序付けは，ウイルスの進化の流れを反映していることがわかる．

4.3　グループとそれぞれのアミノ酸置換の残基位置の関係

グループと各残基位置におけるアミノ酸置換の関係を図 3 に示す．図 3 において，X 軸は幹上のノード ID，Y 軸は各グループにおいて各残基位置 (アイテム) が含まれたトランザクションの数，Z 軸はアミノ酸置換が起こった残基位置，を示す．

図 3 において，各残基位置におけるアミノ酸置換がどの時期においても一定の確率で置換するならば，各グループに対して，各残基位置を含むトランザクションの出現状況は一定であるはずである．つまり，ランダムにどのグループを選択しても，同程度の割合でアミノ酸は置換すると考えられる．しかし，図 3 が示すように，各グループに対してアミノ酸置換の出現状況は一定ではない．例えば，幹の節 ID:254 は多くのトランザクションを含み，もしアミノ酸が一定の割合で

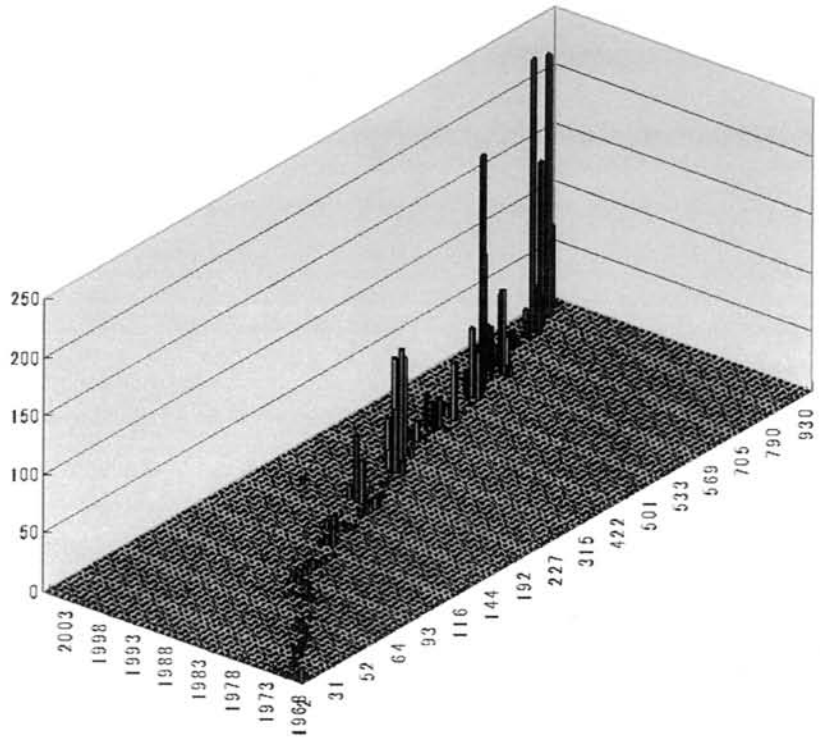


図 2: 幹上の節とウイルス株の年代の関係

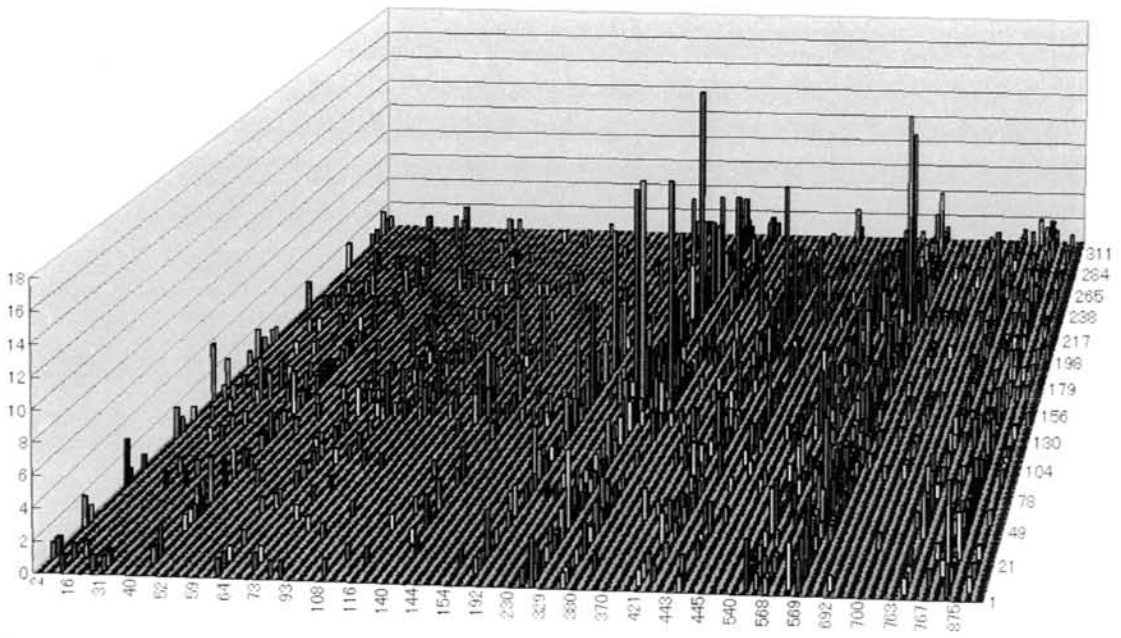


図 3: グループと各残基位置におけるアミノ酸置換の関係

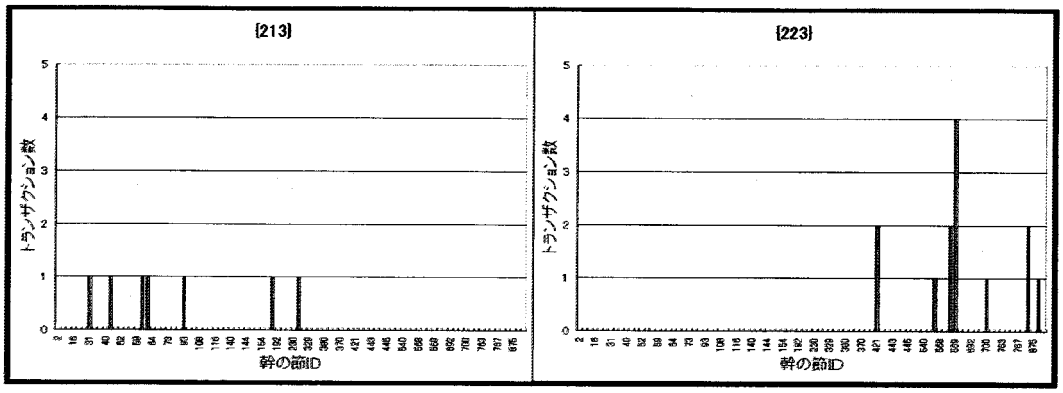


図 4: 本実験で見つかったアミノ酸置換の時代的変化の一例

置換するならば，全体の傾向をよく反映すると考えられるが，全体の状況に依存せずに，138 番や 145 番のようなよく置換する少数の残基位置とあまり置換しないほとんどの残基位置が存在する．このことより，アミノ酸は各グループにおいて常に一定の割合で置換するわけではないことがわかる．

4.4 インフルエンザウイルスのアミノ酸置換における時代的変化

前節まで説明してきたアルゴリズムを C 言語で実装し，1.00 GB RAM，Xeon 3.60 GHz processor のスペックを持つ PC 上で実験を行った．

パラメータ δ を 0.01 に設定し実験を行った結果，アミノ酸置換の頻度が有意に異なる (X, H_k, T_k) を 3129 個発見した．

図 4 に結果の一部を示す．図 4 において，X 軸は幹上のノード ID，Y 軸はトランザクション数を表す．実験結果の可読性を向上させるため図 4 において，トランザクション数を示しているが，実際の計算は本研究における定式化に従い，トランザクションの出現確率を用いて計算している．

図 4 において 213 番のアミノ酸は前半のグループでは置換していたが，後半のグループでは置換しなくなったことがわかる．逆に，223 番のアミノ酸は，前半のグループでは置換していなかったが，後半のグループでは置換するようになったことがわかる．他の約 3000 個の分割 $((X, H_k, T_k))$ においても同様に，アミノ酸置換の分布が均一でない．このような多くの分割が発見されたことにより，各残基位置の置換は各グループに関して均一ではなことが示された．

5 おわりに

本論文では，コントラストセットマイニングの枠組みを用いて，隣接するグループ列間の特徴的違いを発見するアルゴリズムを提案した．本研究の問題には，従来のコントラストセットマイニング問題の研究の成果を直接利用することはできないため，探索空間を効率よく減少させる手法を開発して用いた．インフルエンザウイルスの進化において，アミノ酸置換の起こる残基位置が時代と共に変化するか否かを明らかにするために，進化系統解析と提案手法を組み合わせることによって，アミノ酸置換の起こる残基位置の時代的変化を解析した．その結果，アミノ酸置換の頻度が異なる約 3000 個の特徴的分割 $((X, H_k, T_k))$ が見つかり，インフルエンザウイルスの進化において，各残基位置の置換は各グループに関して均一ではないことが示された．

謝辞

This work was supported, in part, by the Program of Founding Research Centers for Emerging and Reemerging Infectious Diseases from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. We thank Teiji Murakami for his excellent technical assistance.

参考文献

[1] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, In: J. B. Bocca, M. Jarke and C. Zaniolo (Eds.), the 20th Int'l Conf. on Very Large Data Bases, Morgan Kaufmann, VLDB'94, pp. 487–499, 1994.

[2] S. D. Bay and M. J. Pazzani, Detecting Group Differences: Mining Contrast Sets, Data Mining and Knowledge Discovery, Springer Verlag, vol. 5, no. 3, pp. 213–246, 2001.

[3] R. J. Bayardo Jr., Efficiently Mining Long Patterns from Databases. In: L. M. Haas and A. Tiwary (Eds.), the ACM-SIGMOD Int'l Conf. on Management of Data, ACM Press, pp. 85–93, 1998.

[4] G. Dong and J. Li, Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In: the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 43–52, 1999.

[5] S. Hettich, and S. D. Bay, The UCI KDD Archive, Department of Information and Computer Science, University of California, Irvine, CA, http://kdd.ics.uci.edu, 1999.

[6] P. Kralj, N. Lavrac, D. Gamberger and A. Krstacic, Contrast Set Mining Through Subgroup Discovery Applied to Brain Ischaemina Data. In: Zhi-Hua Zhou, H. Li and Q. Yang (Eds.): Advances in Knowledge Discovery and Data Mining, 11th Pacific-Asia Conference, Springer, Proceedings. lnai 4426, PAKDD 2007, pp. 579–586, 2007.

[7] J. Lin and E. J. Keogh, Group SAX: Extending the Notion of Contrast Sets to Time Series and Multimedia Data. In: J. Furnkranz, T. Scheffer and M. Spiliopoulou (Eds.), Knowledge Discovery in Databases, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, Springer, lnai 4213, PKDD 2006, pp. 284–296, 2006.

[8] T. Uno, M. Kiyomi and H. Arimura, LCM ver.2: Efficient Mining Algorithms for Frequent/Closed/Maximal Itemsets. In: R. J. Bayardo Jr., B. Goethals and M. J. Zaki (Eds.), the IEEE International Conference on data mining, 2nd Workshop on Frequent Itemset Mining Implementations (FIMI'04), CEUR-WS.org, CEUR Workshop Proceedings, vol. 126, 2004.

[9] G. I. Webb, S. M. Butler and D. A. Newlands, On detecting differences between groups. In: L. Getoor, T. E. Senator, P. Domingos and C. Faloutsos (Eds.), the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 256–65, 2003.

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

[11] J. Felsenstein. PHYLIP (Phylogeny Inference Package) version 3.66. Department of Genome Sciences, University of Washington, Seattle, 2005.

[12] R. Webster, W. Bean, O. Gorman, T. Chambers, and Y. Kawaoka. Evolution and ecology of influenza A viruses. *Microbiology and Molecular Biology Reviews*, 56(1):152–179, 1992.

[13] D. Wheeler, T. Barrett, D. Benson, S. Bryant, K. Canese, V. Chetvernin, D. Church, M. DiCuccio, R. Edgar, S. Federhen, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 35(Database issue):D5, 2007.

[14] P. Wright and R. Webster. Orthomyxoviruses. *Fields virology*, 1:1533–1579, 2001.