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　 ８ 日現在

機関番号：１７１０４
研究種目：挑戦的萌芽研究
研究期間：2013～2014
課題番号：２５５４０１３７

研究課題名（和文）位置間の進化系統距離に基づくインフルエンザウイルス塩基配列集合の解析

研究課題名（英文）Analyzing Nucleotide Sequences of Influenza Viruses Based on Phylogenetic Distance between Positions

研究代表者

平田　耕一（Hirata, Kouichi）

九州工業大学・大学院情報工学研究院・教授

研究者番号：２０２７４５５８

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：本研究では，塩基配列から進化系統樹を推定し，葉のラベルをある位置の塩基に置換した再ラベル進化系統樹，および，その二分枝を剪定することで得られる剪定進化系統樹に基づいてA型インフルエンザウイルスの塩基配列の解析に取り組んだ．まず，剪定進化系統樹間の距離としての剪定距離を導入し，剪定距離に基づくクラスタリングによってA型インフルエンザウイルスのパンデミックを解析した．次に，進化系統樹に適用可能かつ多項式時間計算可能な合致部分木マッピングカーネルと葉間パスカーネルを導入し，再ラベルおよび剪定進化系統樹を類別用することで，A型インフルエンザのパンデミック，地域間相違，パッケージング位置を解析した．

研究成果の概要（英文）：In this research, we have analyzed nucleotide sequences of influenza A viruses based on the relabeled phylogenetic tree obtained by replacing labels of leaves in a reconstructed phylogenetic tree with nucleotides at a position and the trimmed phylogenetic tree obtained by trimming binary branches in a relabeled phylogenetic tree as possible. First, by introducing a trim distance between trimmed phylogenetic trees, we have applied the clustering to positions based on the trim distance as pandemic analysis for influenza A viruses. Next, by introducing an agreement subtree mapping kernel and a leaf-path kernel that are tractable and applicable to phylogenetic trees, we have classified relabeled and trimmed phylogenetic trees as pandemic analysis, regional analysis and analysis of positions in packaging signals for influenza A viruses.

研究分野：知能情報学

キーワード：進化系統樹　再ラベル進化系統樹　剪定進化系統樹　剪定距離　合致部分木マッピングカーネル　葉間パスカーネル　インフルエンザウイルス

１．研究開始当初の背景

（１）バイオインフォマティクスやメディカルインフォマティクスの分野において，塩基配列集合のどの位置の塩基が進化系統に影響を与えているかを解析することは重要な課題の一つである．この問題に対して，塩基配列集合の位置ごとの塩基の出現確率やエントロピーによって定義される位置間の距離，すなわち，**頻度距離**が最もよく利用されてきた．

（２）ところが，この頻度距離では，いくつかの例外を除いてほとんど同一となる塩基が出現するような塩基配列集合の位置を区別することができない．そこで本研究では，塩基配列集合の位置を特徴づけるために，塩基配列集合から再構成された**進化系統樹**を利用する．

（３）進化系統樹そのものは，生物種の進化を表現する手法として広範囲に用いられている．一方，通常，進化系統樹はただ一つ再構成することが目的であるため，複数の進化系統樹から機械楽手やデータマイニングの手法を適用し知識を発見する手法は，これまでに存在しなかった．

２．研究の目的

（１）そこで本研究では，ただ一つだけ再構成される進化系統樹の葉のラベルを，ある位置におけるそれぞれの塩基配列に出現する塩基に置き換えることで得られる**再ラベル進化系統樹**，および，再ラベル進化系統樹の同一ラベルを持つ葉の二分枝を剪定する，という**ラベルに基づく近隣剪定法**によって得られる**剪定進化系統樹**を新たに導入する．これらの再ラベル進化系統樹および剪定進化系統樹は，塩基配列長と同数得ることができる．

（２）そして本研究では，剪定進化系統樹間の距離となる**剪定距離**を定式化する．この進化系統距離は，塩基配列集合から１本だけに対して定式化される距離である．そして，**剪定距離に基づくクラスタリング**を利用して，インフルエンザ塩基配列の**パンデミック**を解析する．

（３）次に本研究では，再ラベル進化系統樹，および，剪定進化系統樹の類別によって，インフルエンザウイルスを解析する．そのために，進化系統樹に適用可能な**進化系統樹カーネル**を開発し，インフルエンザウイルスの**パンデミック前後や地域間での相違**，および，**パッケージングシグナル位置**を解析する．

３．研究の方法

（１）本研究では，まず，進化系統樹に基づく距離として，２つの剪定進化系統樹間に，二つの木のノードの対応関係である**木マッピング**の変種を利用した**剪定距離**を導入する．最近共通先祖を保存する木マッピングの最小コストとなる距離を **LCA 保存剪定距離**，最大合致部分木(MAST)を保存する木マッピングの最小コストとなる距離を **MAST 剪定距離**として定式化する．

（２）そして，それぞれの距離に基づき，インフルエンザウイルス塩基配列位置の**クラスタリング**を行うことで，2009 年に発生したA 型 H1N1 インフルエンザウイルスのパンデミックを解析する．特に，パンデミック前の 2008 年とパンデミック後の 2009 年の塩基配列に対して，それぞれの塩基配列の位置に基づく剪定距離による**分離位置クラスタリング**，および，2008 年と 2009 年の位置を混在させた剪定距離による**混合位置クラスタリング**，という二つのクラスタリング手法を用いてパンデミックを特徴づける．

（３）次に，再ラベル進化系統樹や剪定進化系統樹に対する**木カーネル**を設計する．まず，二つの木のマッピングを数え上げる**木マッピングカーネル**は，進化系統樹が，葉にしかラベルがなくすべての中間ノードの子の数が 2 となる**無順序葉ラベル全二分木**となるため，その計算は#P 完全となる．そこで本研究では， LCA と葉ラベルを保存し，かつ，対応する内部ノードが必ず二つの葉の LCA となる，すなわち，合致部分木を誘導するような木マッピングを数え上げる**合致部分木マッピングカーネル**を新たに導入する．また，進化系統樹の葉の間のパスを数え上げる**葉間パスカーネル**を新たに導入する．これら 2 つのカーネルを**進化系統カーネル**という．

（４）そして，進化系統カーネルに加えて，比較対象として塩基配列そのものを直接扱う**塩基配列カーネル**を実装し，インフルエンザウイルスのパンデミックや地域間相違，および，パッケージングシグナル位置について解析する．

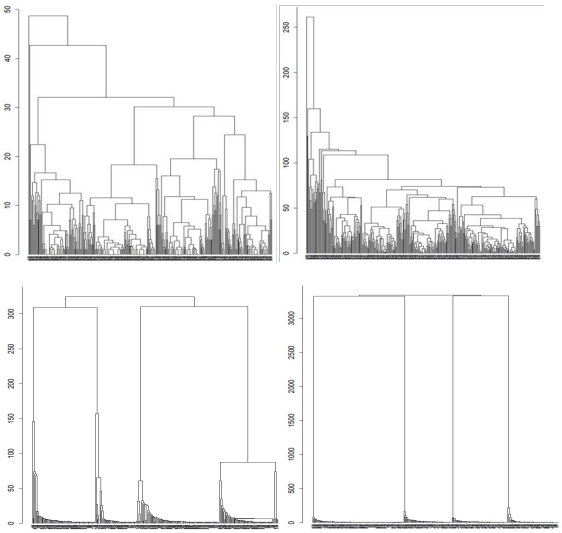
４．研究成果

（１）まず，**分離位置クラスタリング**を A 型 H1N1 亜型インフルエンザウイルスの塩基配列に適用する．ここで，分離位置クラスタリングとは，パンデミック前の 2008 年の塩基配列の位置とパンデミック後の 2009 年の塩基配列位置を分けて考え，それらの位置のうち，同一塩基だけとなる位置を除いて共通な位置について，それぞれでクラスタリングを適用することである．

剪定距離およびハミング距離による分離位置クラスタリングによって，以下の樹系図を得ることができた．上が LCA 保存剪定距離，下が比較対象とするハミング距離，左が 2008 年，右が 2009 年の分離位置クラスタリングの結果である．

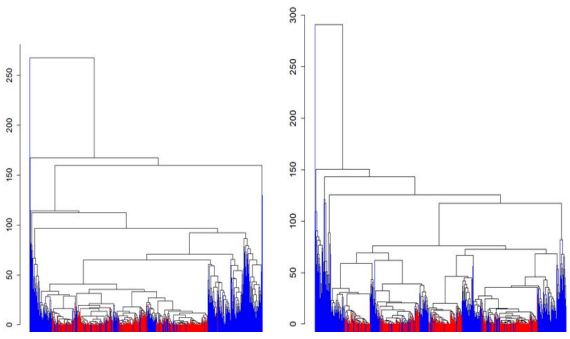
ここで，LCA 保存剪定距離の樹系図の高さ

は，2008 年が 48，2009 年が 261 であり，ハミング距離の樹系図の高さは，2008 年が 324，2009 年が 3340 である．ハミング距離では樹系図の高さに対してクラスタが低い位置に集中しているため分解能が非常に低くなるのに対して，LCA 保存剪定距離では，有意なクラスタができていることが分かる．MAST 剪定距離でも，LCA 保存剪定距離と同様の傾向となった．



（２）次に，**混合位置クラスタリング**を A 型 H1N1 亜型インフルエンザウイルスの塩基配列に適用する．ここで，混合位置クラスタリングとは，パンデミック前の 2008 年の塩基配列の位置とパンデミック後の 2009 年の塩基配列位置を合わせて同時にクラスタリングを適用することである．ここで， 2008 年と 2009 年の塩基配列数が異なるため，この手法をハミング距離に適用することはできない．

剪定距離による混合位置クラスタリングの結果，以下の樹系図を得ることができた．ここで，左が LCA 保存剪定距離，右が MAST 剪定距離による混合位置クラスタリングの結果であり，樹系図の高さはそれぞれ 267，291 である．また，赤が 2008 年，青が 2009 年の塩基配列を表している．



どちらの剪定距離からも類似した樹系図が得られた．また，これらの樹系図から，2008 年と 2009 年はそれぞれクラスタが固まる傾向にあった．ただし，厳密に 2008 年と 2009 年のデータを類別するクラスタは見つからなかった．

これらのクラスタリングの結果により，パンデミック前後は，別の類別手法を用いることでより厳密な結果が期待できることが分かった．そこで次に，クラスタリングよりも直接塩基配列やその位置を類別する手法としての**進化系統樹カーネル**について研究を進めた．

（３）まず，**木マッピングカーネル**の一つとして，合致部分木を誘導するような木マッピングを数え上げる**合致部分木マッピングカーネル**を新たに導入し，無順序葉ラベル全二分木である進化系統樹に対して，それを葉の数の二乗時間で計算するアルゴリズムを設計した．一方，各内部ノードの子の数が制限されていない無順序葉ラベル木の場合には，合致部分木マッピングカーネルの計算は#P 完全になることを示した．

また，進化系統樹の部分構造を数え上げるカーネルとして，進化系統樹の葉の間のパスを数え上げる**葉間パスカーネル**を新たに導入し，それを葉の数の線形時間で計算する亜リズムを設計した．

（４）これらの進化系統樹カーネルに加えて，比較対象として塩基配列そのものを直接扱う配列カーネル，重集合カーネル，文字列カーネルを**塩基配列カーネル**として実装し，A 型 H1N1 亜型および H3N2 亜型インフルエンザウイルス塩基配列を解析した．

以下の表は，表の上に示されたそれぞれのカーネルに基づき SVM を用いて類別した結果の概要である．

	亜型	塩基配列	合致部分木	葉間パス
パンデミック解析	H1N1	高精度	高精度	完全
地域間相違解析	H1N1 H3N2	低精度	高精度	完全
パッケージングシグナル位置解析	H1N1 H3N2	高精度	低精度	高精度

ここで，***パンデミック解析***では，パンデミック前の 2008 年の塩基配列を負例，パンデミック後の 2009 年の塩基配列として類別している．また，**地域間相違解析**では，ある大陸の塩基配列を正例，別の大陸の塩基配列を負例として，すべての大陸間の組合せにおいて類別している．

一方，***パッケージングシグナル***とは，インフルエンザウイルスの粒子化において 8 つの RNA 分節が選択的に集合するための塩基のことであり，これらの位置はリバースジェネテ

ィックスにおいて解析されている．そこで，パッケージングシグナル位置解析では，パッケージングシグナル位置を正例，そうでない位置を負例として扱うことで類別している．

（５）上の表により，**パンデミック解析**では，どのカーネルでも高精度で類別することができた．また，**地域間相違解析**において塩基配列カーネルによる類別は低精度であったのに対して，進化系統樹カーネルによる類別は高精度であり，特に，葉間パスカーネルでは完全に類別することができた．一方，**パッケージングシグナル位置解析**では，合致部分木マッピングカーネルによる類別は低精度であったのに対して，塩基配列カーネルと葉間パスカーネルは高精度で分類することができた．その結果，本研究で扱ったパンデミック解析，地域間相違解析，パッケージングシグナル位置解析では，**葉間パスカーネル**による類別が最も成功した．

（６）一方，本研究では，カーネルによる類別結果を利用することで，塩基配列におけるそれぞれの問題の原因究明，という点にまでは至っていない．特に，パッケージングシグナル位置解析では，8 つの RNA 分節が選択的に集合する塩基であることから，RNA 配列の二次構造や三次結合の解析を本研究の成果と融合して新たな知見を得ることは，今後の重要な課題である．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　６件）

①Issei Hamada, Takaharu Shimada, Daiki Nakata, Kouichi Hirata, Tetsuji Kuboyama : *Classifying Nucleotide Sequences and Their Positions of Influenza A Viruses through Several Kernels*, Proc. 4th International Conference on Pattern Recognition Applications and Methods (ICPRAM2015), 342-347, 2015 (査読有)
DOI 10.5220/0005251103420347

②Issei Hamada, Takaharu Shimada, Daiki Nakata, Kouichi Hirata, Tetsuji Kuboyama : *Agreement Subtree Mapping Kernel for Phylogenetic Trees*, New Frontier in Artificial Intelligence, Lecture Notes in Artificial Intelligence 8417, 1--16. 2014 (査読有)
DOI 10.1007/978-3-319-10061-6_2

③Issei Hamada, Kouichi Hirata, Tetsuji Kuboyama, Takaharu Shimada : *Agreement-Subtree Mapping Kernel and Leaf-Path Kernel for Phylogenetic Trees Reconstructed from Nucleotide Sequences*, Proc. Workshop on Graph-Based Algorithms for Big Data and Its Application (GABA2014), 2014 (査読有)

④Keisuke Ueno, Akihiro Ishii, Kimihito Ito : *ELM: Enhanced Lowest Common Ancestor Based Method for Detecting a Pathogenic Virus from a Large Sequence Dataset*, BMC Bioinformatics 15. 254, 2014 (査読有)
DOI 10.1186/1471-2105-15-254

⑤伊藤公人 : *計算機科学によるインフルエンザウイルスの抗原変異予測*, 化学療法の領域 30, 68-74, 2014 (査読有)

⑥Takaharu Shimada, Issei Hamada, Kouichi Hirata, Tetsuji Kuboyama, Kouki Yonezawa, Kimihito Ito : *Clustering of Positions in Nucleotide Sequences by Trim Distance*, Proc. IIAI International Conference on Advanced Applied Informatics (IIAI AAI 2013), 129-134, 2013 (査読有)
DOI 10.1109/IIAI-AAI.2013.72

〔学会発表〕（計　９件）

①Kimihito Ito : *Prediction of Amino Acid Substitutions on the Hemagglutinin Molecules of H3N2 Seasonal Influenza Viruses Using Gamma Distribution*, The ARI conference in The U.S-Japan's Cooperative Medical Sciences Program, 2015 年 1 月 28 日, Academia Sinica (Taipei, Taiwan)

②伊藤公人 : *コンピューターは季節性インフルエンザウイルスの抗原変異を予測できるか？*, 第 15 回 IPAB シンポジウム／感染症とコンピュータ創薬, 2014 年 12 月 5 日, NEC 本社(東京都港区)

③Kimihito Ito : *Predicting the Evolution of Influenza A Viruses through Data Assimilation* (Keynote Lecture), International Workshop on Information Search, Integration & Personalization (ISIP2014), 2014 年 10 月 9 日, Help University (Kuala Lumpur, Malaysia)

④伊藤公人 : *コンピューターでインフルエンザウイルスの変異を予測する*, 第 157 回日本獣医学会学術集会 特別企画「人獣共通感染症の先回り対策策定に向けた情報戦略」, 2014 年 9 月 11 日, 北海道大学(北海道札幌市)

⑤濱田一青, 島田昴治, 中田大貴, 平田耕一 : *さまざまなカーネルによる A 型インフルエンザウイルスの塩基配列解析*, 人工知能学会基本問題研究会(第 94 回), 人工知能学会

研究会資料 SIG-FPAI-B401, 1—6, 2014 年 7 月 24 日, 根室市総合文化会館(北海道根室市)

⑥中田大貴, 濱田一青, 島田昴治, 平田耕一 : *塩基配列へのカーネルの適用による A 型 H1N1 インフルエンザウイルスの地域解析*, 人工知能学会基本問題研究会(第 92 回), 人工知能学会研究会資料 SIG-FPAI-B303, 47—52, 2014 年 1 月 30 日—31 日, 函館市民会館(北海道函館市)

⑦島田昴治, 濱田一青, 平田耕一 : *進化系統樹の葉間パスカーネルによるパンデミック解析*, 人工知能学会基本問題研究会(第 91 回), 人工知能学会研究会資料 SIG-FPAI-B302, 7–12, 2013 年 11 月 28 日—29 日, 愛媛大学(愛媛県松山市)

⑧濱田一青, 島田昴治, 平田耕一 : *進化系統樹の合致部分木マッピングカーネルによるパッケージングシグナル位置解析*, 人工知能学会基本問題研究会(第 91 回), 人工知能学会研究会資料 SIG-FPAI-B302, 1–6, 2013 年 11 月 28 日—29 日, 愛媛大学(愛媛県松山市)

⑨島田昴治, 濱田一青, 平田耕一, 久保山哲二, 米澤弘毅, 伊藤公人 : *剪定距離による塩基配列位置のクラスタリング*, 夏の LA シンポジウム 2013, 2013 年 7 月 16 日—18 日, 休暇村志賀島(福岡県福岡市)

〔その他〕
ホームページ
http://www.dumbo.ai.kyutech.ac.jp/hirata

６．研究組織
(1)研究代表者
平田　耕一（HIRATA, Kouichi）
九州工業大学・大学院情報工学研究院・教授
研究者番号：20274558

(2)研究分担者
伊藤　公人（ITO, Kimihito）
北海道大学・人獣共通感染症リサーチセンター・教授
研究者番号： 60396314