科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　１９　日現在

機関番号：　３４２０４
研究種目：　若手研究(B)
研究期間：　2012～2013
課題番号：　２４７００２８９

研究課題名（和文）系統樹の剪定による遺伝子配列データリサンプリングアルゴリズム

研究課題名（英文）The Closest-Neighbor Trimming Algorithm for Resampling Genetic Sequence Datasets

研究代表者
米澤　弘毅（Yonezawa, Kouki）

長浜バイオ大学・バイオサイエンス学部・助手

研究者番号：００３７４７４４

交付決定額（研究期間全体）：（直接経費）　　3,300,000円、（間接経費）　　990,000円

研究成果の概要（和文）：近年，インフルエンザを始め様々な病原体の遺伝子情報がデータベース上に大量に蓄積されている．しかし，データセットの巨大化に伴い，多重配列アラインメント，進化系統解析，相同性検索等の解析が困難になりつつある．また，各国における感染症サーベイランス能力の差異により，データセットにサンプリングバイアスを含んでいる可能性がある．
本研究では，サンプリング密度の濃い部分に存在する配列を適宜取り除くことによってリサンプリングを行うアルゴリズムを提案し，インフルエンザウイルスなどの病原体の遺伝子配列データに適用することにより，サンプリングバイアスの軽減能力を評価した．

研究成果の概要（英文）：A large number of nucleotide sequences of various pathogens are available in public databases. The growth of the datasets has resulted in an enormous increase in computational costs. Moreover, due to differences in surveillance activities, the number of sequences found in databases varies from one country to another and from year to year. Therefore it is important to study resampling methods to reduce the sampling bias. A novel algorithm, called the closest-neighbor trimming method, that resamples a given number of sequences from a large nucleotide sequence dataset was proposed. The performance of the proposed algorithm was compared with other algorithms by using the nucleotide sequences of human H3N2 influenza viruses. Since nucleotide sequences are among the most widely used materials for life sciences, we anticipate that our algorithm to various datasets will result in reducing sampling bias.

研究分野：　情報学

科研費の分科・細目：　情報学フロンティア　生命・健康・医療情報学

キーワード：　バイオインフォマティクス　人獣共通感染症　リサンプリング　分子系統樹

１．研究開始当初の背景

近年，インフルエンザを始め様々な病原体の遺伝子情報がデータベース上に大量に蓄積されている．例えば 2011 年 10 月 6 日現在，173,595 件の塩基配列が NCBI Influenza Virus Resource に存在する．このような大量のデータセットは病原体の研究者にとっては有用に思われるが，多重配列アラインメント，進化系統解析，相同性検索等の解析を行うには膨大な計算時間を要する．また，公開されているデータベースにはサンプリングバイアスを伴う危険性があることが指摘されている．例えば NCBI Influenza Virus Resource に存在するヒトの H3N2 インフルエンザウイルスにおいて，1968 年の香港カゼの流行以降のデータが蓄積されているが，PCR 法が発明された 1982 年以降登録される配列数が増加し始め，実に 92%以上が 1992 年以降に分離されたデータである(図 1)．また，データが登録された国の分布を見ると，30%以上がアメリカからの登録であることがわかっている．この偏りは各国のサーベイランス能力に差異があることが主な要因であろうと考えられる．また，NCBI Influenza Virus Resource に登録されている A 型インフルエンザウイルスの全 8 種のセグメントを見ても，全てのセグメントにおいて 7,000 配列以上のデータを有し，少なくとも約半分はヒトから分離されたデータである．データベース内に存在する遺伝子情報データには明らかにバイアスが含まれており，後に続く解析に支障をきたす危険性を含んでいることがわかる．

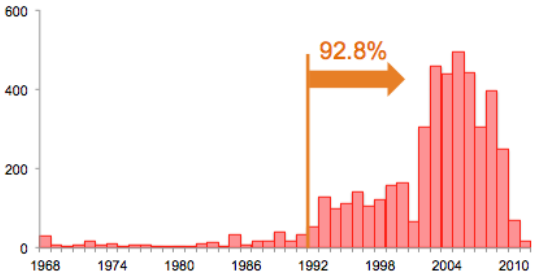

図 1: ヒト H3N2 インフルエンザウイルスの分離年の分布. 1992 年以降のデータが 92.8%を占める．

２．研究の目的

本研究では，サンプリング密度の濃い部分に存在する配列を適宜取り除くことによってリサンプリングを行うアルゴリズムを提案する．本手法を様々な病原体の遺伝子配列データに適用することにより，サンプリングバイアスの軽減能力を評価する．具体的には，(1)系統樹の剪定による配列データリサンプリングアルゴリズムの開発，(2)リサンプリングアルゴリズムに対する評価手法の確立，および(3)様々な病原体へのアルゴリズムの応用の 3 点の達成を目的とする．

３．研究の方法

申請者は研究期間内に，(1)系統樹の剪定による配列データリサンプリングアルゴリズムの開発，(2)リサンプリングアルゴリズムに対する評価手法の確立，および(3)様々な病原体へのアルゴリズムの応用を実行する．(1)系統樹剪定アルゴリズムの開発においては，系統樹において剪定する配列の選び方について吟味し，また従来より提案されてきた距離行列や配列情報そのものを用いて行うサンプリングの方法との性能比較を行う．(2)剪定アルゴリズムによって配列が適切に剪定されているか，また出力された系統樹と従来から得られている知見が合致するかどうかで評価を行う．(3)インフルエンザウイルス，C 型肝炎ウイルス，サル免疫不全ウイルスなど，配列情報がデータベースに豊富に存在する病原体に対して剪定アルゴリズムを適用し，結果について考察を行う．

４．研究成果

本研究では，サンプリング密度の濃い部分に存在する配列を適宜取り除くことによってリサンプリングを行うアルゴリズムを提案した．申請者らは塩基多型度(図 2)やサンプリングバイアスの軽減能力(図 3)を他のアルゴリズムと比較してその優位性を示した．また，設計したリサンプリングアルゴリズムをウェブ上に公開した．

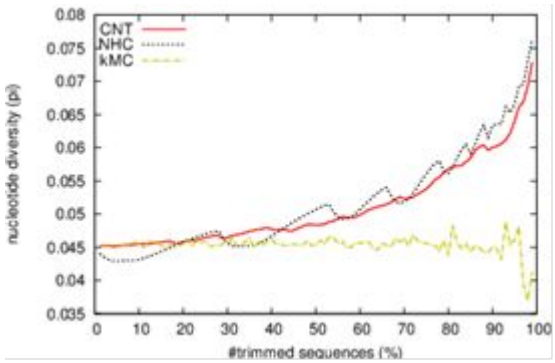

図 2: リサンプリングアルゴリズムを適用して得られるデータセットの配列数と塩基多型度の関係．図中の赤色の折れ線が申請者らのアルゴリズムの性能を示す．

さらに，申請者らは A 型インフルエンザウイルスに対して本手法を適用し，結果の考察を行った．インフルエンザウイルスが持っている全 8 セグメントの遺伝子情報を用いて系統樹を作成した後に本手法を適用し，過去に得られた知見を参照しつつ得られた系統樹に対する個別の議論を行った．特に 2 万以上の配列数を持つセグメントの場合，元のデータセットから作成した系統樹とリサンプリングしたコンパクトなデータセットから作成した系統樹を比較すると，その親子関係が差異が生じるケースが確認された．

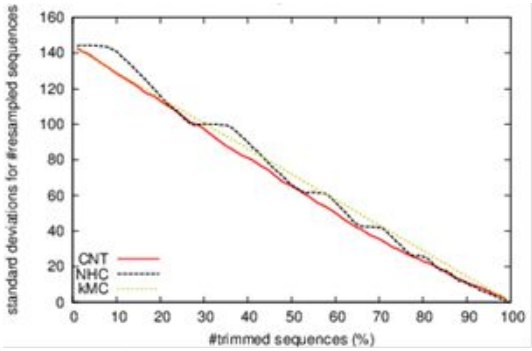

図 3: 1 年に投稿される配列数の標準偏差とリサンプリングによって得られる配列数との関係．標準偏差が小さいほど分離年による偏りが小さい，すなわちサンプリングバイアスが小さいことを示唆する．

さらに，A 型インフルエンザウイルスのうち鳥を宿主とするものの塩基配列や他の病原体ウイルスの塩基配列についても申請者らのアルゴリズムを適用し，実験を行う研究者らにコンパクトなデータセットを提供することが出来た．また，Web サービスにおいて申請者らのリサンプリングアルゴリズムを公開した．本サービスにおいては，ユーザがリサンプルしたい配列データセット(塩基配列とアミノ酸配列の両者に対応)とリサンプルするサイズを入力し，近隣剪定法によってリサンプルされた結果の URL を記述したメールを受け取り，データセットのダウンロードを行う．

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 3 件）

1. Kouki Yonezawa, Manabu Igarashi, Keisuke Ueno, Ayato Takada, Kimihito Ito. Resampling nucleotide sequences with closest-neighbor trimming and its comparison to other methods. PLoS One, 8(2), 2013, e57684, 10.1371/journal.pone.0057684

2. Shunsuke Makino, Takaharu Shimada, Kouichi Hirata, Kouki Yonezawa, Kimihito Ito. A Trim Distance between Positions as Packaging Signals in H3N2 Influenza Viruses. The 6th International Conference on Soft Computing and Intelligent Systems and the 13th International Symposium on Advanced Intelligent Systems, 2012.

3. Shunsuke Makino, Takaharu Shimada, Kouichi Hirata, Kouki Yonezawa, Kimihito Ito. A Trim Distance between Positions as Packaging Signals in H3N2 Influenza Viruses. The 6th International Conference on Soft Computing and Intelligent Systems and the 13th International Symposium on Advanced Intelligent Systems, 2012.

〔学会発表〕（計 7 件）

1. Kouki Yonezawa, Manabu Igarashi, Kimihito Ito. Resampling of Large Nucleotide Sequence Datasets of Viruses with Closest-Neighbor Trimming Method. Annual Meeting of Society for Molecular Biology & Evolution (SMBE 2013), 2013 年 7 月 7 日-11 日.

2. 米澤 弘毅, 五十嵐 学, 伊藤 公人. 近隣剪定法によるウイルス遺伝子配列のリサンプリング. 第 7 回ゲノム微生物学会年会, 長浜バイオ大学 , 2013 年 3 月 8 日-10 日.

3. Takaharu Shimada, Shunsuke Makino, Kouichi Hirata, Kouki Yonezawa, Kimihito Ito. Clustering of Positions in Nucleotide Sequences by Trim Distance. CSBB 2013, のがみプレジデントホテル, 2013 年 2 月 28 日-3 月 1 日.

4. 米澤 弘毅, 五十嵐 学, 伊藤 公人. 近隣剪定法によるウイルス遺伝子配列のリサンプリング. 感染症若手フォーラム 2013, 北広島クラッセホテル, 2013 年 2 月 28 日-3 月 2 日.

5. 米澤 弘毅, 五十嵐 学, 伊藤 公人. 近隣結合法による塩基配列データセットのリサンプリングとその性能評価. 第 88 回人工知能基本問題研究会, 石垣市民会館, 2013 年 1 月 24 日-25 日.

6. 米澤 弘毅, 五十嵐 学, 伊藤 公人. 近隣剪定法：進化系統樹を利用した配列リサンプリングアルゴリズム. 第 60 回日本ウイルス学会学術集会, 大阪国際会議場, 2012 年 11 月 13 日-15 日.

7. 米澤 弘毅, 五十嵐 学, 伊藤 公人. 近隣剪定法: 進化系統樹を利用した配列リサンプリングアルゴリズム. 第 29 回バイオ情報学研究会, 沖縄科学技術大学院大学, 2012 年 6 月 28 日-29 日.

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ等
Biological Sequence Resampling

http://citrus.nagahama-i-bio.ac.jp/resampling/

６．研究組織

(1)研究代表者

米澤 弘毅　(Yonezawa, Kouki)
長浜バイオ大学・バイオサイエンス学部・助手
研究者番号：00374744

(2)研究分担者

(　　)

研究者番号：

(3)連携研究者

(　　)

研究者番号：