Vol.2012-BIO-29 No.22
2012/6/29

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

米澤弘毅[†]　五十嵐学[††]　伊藤公人[††]

近年，インフルエンザをはじめ様々な病原体の遺伝子情報が大量に蓄積されつつある．データセットの増大に伴い，配列解析にかかる計算コストが急増している．また，疫学調査活動の差異により，データセットは調査地域や年代に関して大きなサンプリングバイアスを含む．本研究では，進化系統樹を利用してサンプリング密度の高い配列を適宜取り除くリサンプリングアルゴリズムを提案し，その性能を比較実験により評価する．

The closest-neighbor trimming method: A resampling algorithm for nucleotide sequence datasets

KOUKI YONEZAWA[†]　MANABU IGARASHI[††]
KIMIHITO ITO[††]

Recently 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. In this paper we propose a novel algorithm–called the closest-neighbor trimming method–that resamples a given number of sequences from a large nucleotide sequence dataset. We compare the performance of the proposed algorithm with other algorithms by using the nucleotide sequences of human H3N2 influenza viruses.

1. はじめに

感染症研究において，病原体の遺伝子情報は，伝播経路の推定，病原性の分子基盤の解明，予防・診断・治療薬の開発に必須な情報となりつつある．国内外の研究機関は，臨床分離検体を収集し，自国で流行する病原体の遺伝子情報を調査している．感染症に国境はないので，病原体の遺伝子情報は共有される必要があり，GenBank 等の公的データベースには,病原体の塩基配列が大量に蓄積されている．例えば，NCBI Influenza Virus Resource [1]には，インフルエンザウイルスの 170,000 本以上の塩基配列が，また HIV sequence database [2]にはヒト免疫不全ウイルスの 410,000 本以上の塩基配列がそれぞれ登録されている．

データベース上の登録配列数の急速な増加は，以下に挙げる二つの問題を引き起こしている．一つは計算コストの増大である．多重配列アラインメント，系統解析，ホモロジー検索といった配列データ解析を大規模に実行するには多大な計算量を必要とする．多重配列アラインメントは NP-完全問題であり[3]，進化系統解析は計算の比較的速い近隣結合法[4]を用いたとしても，n本の塩基配列を扱うのに$O(n^3)$の計算時間がかかる．また，BLAST を用いたホモロジー検索[5]においては，n本の塩基配列およびクエリの部分文字列の長さwに対して$O(wn\log n)$の計算時間がかかる．

もう一つの問題は，公的データベースに登録されている遺伝子データに関するサンプリングバイアスの問題である．サンプリングバイアスは，遺伝子配列が読み取られた株の分布が，母集団，つまり実際の流行株の分布を代表していない場合に発生する．疫学調査活動の規模は国ごとに大きく異なり，積極的に調査を行って多くの塩基配列をデータベースに登録している国もあれば，ほとんど配列データがない国もある．また，ここ 20 年間におけるシークエンシング技術の進歩も別のサンプリングバイアスの要因として挙げられる．一般に，古い流行株については，流行当時の技術的制約から解読されている塩基配列の数が，最近の流行株のそれに比べて圧倒的に少ない．つまり，データベースに登録されている病原体の塩基配列の数は，実際の流行の規模を反映していない．

本研究では，データベースに登録されている塩基配列を適切にリサンプリングする問題を扱う．Zaslavsky らは，進化系統樹から特徴的な枝を残して他を削除することにより，限られた画面領域内に巨大な進化系統樹を近似的に描画するアルゴリズムを提案した[6]．また，クラスタリングアルゴリズムのうち，階層的クラスタリング[7]や*k*-medoids アルゴリズム[8]といった，データポイントをクラスタの代表点とするアルゴリズムもリサンプリングアルゴリズムとして利用できると考えられる．

† 長浜バイオ大学バイオサイエンス学部コンピュータバイオサイエンス学科
Department of Computer Bioscience, Nagahama Institute of Bio-Science and Technology.
†† 北海道大学人獣共通感染症リサーチセンター
Hokkaido University Research Center for Zoonosis Control

本論文では，大きな配列データセットから与えられた数の配列を残してリサンプリングするための新規アルゴリズムを提案する．密にサンプリングされた配列からより多くの配列を除去し，密でない配列をできるだけ多く保持するようにリサンプリングを行えば，データセットに含まれるサンプリングバイアスを減らすことができるはずであり，この基本的アイデアに従ってアルゴリズムを設計した．提案手法を**近隣剪定法**(Closest-Neighbor Trimming Method)と呼び，以下 CNT と略記する．CNT は全配列データセットから進化系統樹を作成し，近隣となるすべての配列ペアのうちで最も距離の短い配列ペアを探索し，そのペアから共通祖先との距離が遠いほうの配列を除去する．この手続きを繰り返すことにより，CNT は密にサンプルされた配列からより多くの配列を除去する．

本論文では CNT と既存のアルゴリズムとの性能評価をヒト H3N2 インフルエンザウイルスのヘマグルチニン(以下 HA と表記する)の塩基配列データを用いて行い，CNT が効率的にサンプリングバイアスを除去していることを示す．

2. 手法およびデータセット

2.1 リサンプリング問題

アルゴリズムについて述べる前に，まずリサンプリング問題を定義する．n本の配列データからなる集合Sに対し，その中からk本の配列からなる部分集合$R \subset S$を選び出すことをリサンプリングという．仮定として，集合S内のすべての配列は同じ長さを有している(もしくはアラインメントされている)．また集合Sはランダムにサンプリングされていない可能性があることに注意する．幾つかの特徴を除いては，元の集合Sに属する配列の特徴は未知である．リサンプリングの目的は，元の集合S内の配列の特徴を反映した部分集合Rを見つけることである．

2.2 近隣剪定法

本論文で我々は密にサンプルされた配列を効率的に除去する CNT というアルゴリズムを提案する．最初に，データセットにあるすべての配列を用いて系統樹を作成する．ここで系統樹の作成法については特に仮定はしない．もし作成した系統樹が二分木でなければ，CNT は任意に二分木に変換する．この二分木を$G = (V, E)$と定義する．ここでVおよびEはそれぞれ接点及び枝の集合を表す．Gが与えられた時，CNT は以下の手続きを残った配列数が与えられたkになるまで繰り返す．系統樹に存在する全ての近隣のペアから最も距離の短いペアを見つけ出し，そのペアのうち共通祖先からの距離が遠いものを除去する．そして除去した後の系統樹$G' = (V', E')$が二分木になるように枝を張り替える．CNT の擬似コードを図 1 に示す．

0. Set the subset of leaves $V' \leftarrow V$,
   the subset of edges $E' \leftarrow E$, and
   the set of edge lengths $\ell' \leftarrow \ell$.
1. While $|V'| > n-m$ do the following:
   1-1. Find a pair of neighbors v, w and their common parent u
   with the shortest distance among all the pair of neighbors, that is,
   $\ell'(u,v)+\ell'(u,w) = \min_{(u',v'),(u',w') \in E'} \{\ell'(u',v')+\ell'(u',w')\}$.
   1-2. If v is further from u than w, then $V' \leftarrow V' / \{v\}$.
   Otherwise $V' \leftarrow V' / \{w\}$.
   Suppose that v is removed.
   1-3. Let u_{rev} be another node connected with u.
   Modify V', E', and ℓ' so that the resulting tree $G' = (V', E')$ is binary, that is,
   $\ell'(u_{rev}, w) = \ell'(u_{rev}, u) + \ell'(u, w)$, $E' \leftarrow E' \cup \{(u_{rev}, w)\} / \{(u_{rev}, u), (u, w)\}$, and $V' \leftarrow V' / \{u\}$.
2. Return V' as the result of resampling.

図 1　CNT の擬似コード

Figure 1　Pseudo-code of CNT.

2.3 比較するリサンプリングアルゴリズム

(1) **Zaslavsky らの手法 (ZAS05)**

Zaslavsky らが提案したリサンプリングアルゴリズムは以下のものである[6]．系統樹Gが与えられると，ZAS05 は最初に 2 本の配列を選び出す．1 本は系統樹の根から最も近い配列，もう 1 本は根から最も遠い配列である．Rを既に選択された配列の集合とし，ある配列sとRとの距離を，sとR内の配列との距離の最小値と定義する．ZAS05 はRへの距離が最も遠い配列を選び出す．この手続きを選択された配列数がkに達するまで繰り返す．

(2) **単純な階層的クラスタリングアルゴリズム (NHC)**

単純な階層的クラスタリングアルゴリズム[7]は実際のデータをクラスタの代表として扱う．つまり，クラスタの代表以外の点を除去することで NHC をリサンプリングアルゴリズムとみなすことができる．$n \times n$の距離行列Dが与えられた時, NHC はすべての配列のペアから最も距離の近いペアを見つけ出し，そのうちその他の配列への距離が遠い方を除去する．この手続を残った配列数がkになるまで繰り返す．

(3) ***k*-medoids クラスタリングアルゴリズム (kMC)**

k-medoids アルゴリズム[8]も配列リサンプリングに適用することができる．$n \times n$の距離行列Dが与えられた時，kMC はランダムにk本の配列を選択する．その後 kMC は，まず各配列を最も近いクラスタに割り当て，そしてクラスタごとに，そのクラスタ内にある他の配列との距離が最小となるように代表すなわち medoids を選出する．この手続きを，medoids が変化しなくなるか繰り返し回数がある閾値に達するまで繰り返す．(本研究ではこの閾値を 1,000 回に設定している)

2.4 データセット

各アルゴリズムの性能評価に用いたデータセットは, NCBI Influenza Virus Resource [1]からダウンロードしたヒト H3N2 インフルエンザウイルスの HA の配列である．この配列を MAFFT [9]でアラインメントした後，短い配列や曖昧な塩基を持つ配列を取り除いた．その結果，984 塩基か

らなる配列 4,655 本を得ることができた.

2.5 性能評価の方法

(1) 系統樹の作成法

CNT と ZAS05 は系統樹を入力とするが，その作成法は問わない．本論文では PHYLIP[10]の近隣結合法[4]を適用した．また，距離行列作成の際には Jukes-Cantor モデル[11]を使用した.

(2) 結果の評価基準

リサンプリング結果を評価するために 2 種類の指標を用いる．1 つ目は$(n-k)$本の除去された配列とk本の残った配列の一致度の平均値である．定義は以下のようになる．まず, 2 本の配列s_1とs_2間の異なる塩基数を$\mathrm{diff}(s_1, s_2)$と表す. これら 2 本の配列間の一致度$I(s_1, s_2)$は 2 本の配列間の同一の塩基数の比で表される．つまり，$I(s_1, s_2) = 1 - \mathrm{diff}(s_1, s_2)/\mathrm{length}(s_1)$と書ける. すると除去された配列の集合$D$とリサンプリングされた配列の集合$R$の間の一致度は以下のように定義できる.

$$I(D;R) = \frac{1}{|D|}\sum_{s'\in R}\max_{s\in R} I(s,s')$$

2 つ目の評価基準として，データセットに存在する配列の分離年の分布を用いる．理想的なデータセットにおいては，ある生物種の塩基配列数はその個体数に比例するはずであるが, 感染症の病原体の実数を知ることは困難である. H3N2 インフルエンザの感染者数は年によって変動するが, データベースに投稿される配列数の年による変動は感染者数の変動よりも大きくなっている. この事実と単純さから, 本論文では配列のデータセットは毎年同じ数の配列を保持しているべきであるという仮定を導入する. この仮定から, 本論文ではサンプリングバイアス軽減能力の評価において, 分離年の配列数に関する標準偏差を用いることとする.

3. 結果

3.1 データセットの分離年および分離国の分布

データセットは 1968 年から 2011 年までに分離されたヒト H3N2 インフルエンザウイルスの HA の塩基配列からなる．そのうち 1968 年から 1991 年までのデータが 7%で, 1992 年以降のデータが約 93%を占める(図 2).

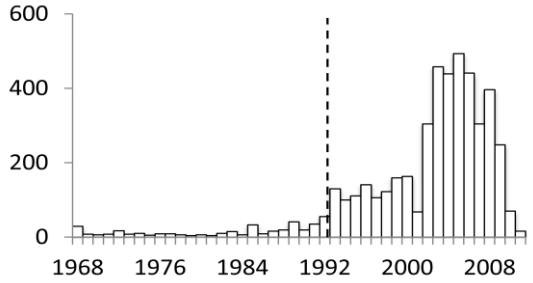

図 2　データセットにおける配列の分離年の分布

Figure 2　Distribution of isolation years of the original dataset.

この歪んだ分布は 1992 年前後のシークエンシング技術の急速な発展によるサンプリングバイアスであろうと考えられる[12]．加えて，30%以上の配列が USA から投稿されたものであり(図 3), これはアメリカの高いサーベイランスがもたらしたサンプリングバイアスが関係するであろうと推測される.

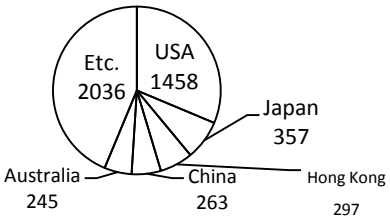

図 3　データセットにおける分離国の分布

Figure 3　Distribution of isolated countries in the dataset.

3.2 リサンプルされた配列の評価

最初に除去された配列とリサンプルされた配列間の一致度$I(D;R)$に関する評価を行う．密にサンプルされた配列を除去できている場合，$I(D;R)$は 100%に近い値を示す. また，このデータセットにおける一致度の最小値が 82.3%であることから，$I(D;R)$は 82.3%を下回ることはない. 図 4 は各アルゴリズムを適用した場合の，除去した配列の割合と$I(D;R)$の関係を示したものである．横軸は除去した配列の割合(%)を，縦軸は$I(D;R)$の値(%)を表す．リサンプリングアルゴリズムのうち，CNT，NHC および kMC は 90%の配列が除去されるまで 100%近くの一致度を保っていた．そこで次の性能評価においては，CNT，NHC および kMC アルゴリズムを対象とする.

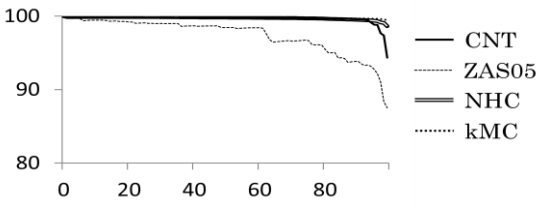

図 3　リサンプリングによる$I(D;R)$の変化

Figure 3　Change of $I(D;R)$ according to the resampling.

次にリサンプルされた配列の分離年の分布について確認を行う．図 4 は CNT，NHC および kMC でリサンプリングを行った際の分離年の分布の変化をヒストグラムで表したものである.

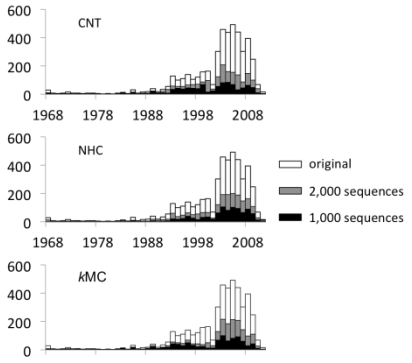

図 4　リサンプリングによる分離年の分布の変化

Figure 4　Change of the distributions of isolation years according to the resampling.

ヒト H3N2 インフルエンザウイルスのデータセットでは, 1 年に投稿される配列数の平均が約 10 本, その標準偏差は約 142 であった．この標準偏差がリサンプリングとともにどう減少するかを見たものが図 5 である．

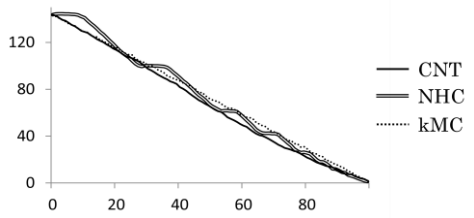

図 5　分離年の標準偏差の変化

Figure 5　Changes of the standard deviations of isolation years.

図 5 において，横軸は除去した配列の割合(%)を，縦軸は分離年の標準偏差を表す．前述の理由により，密にサンプルされた配列を除去できているほど，標準偏差の値は小さくなる．図 5 より，CNT が他のアルゴリズムに比べて標準偏差が小さくなっていることがわかる．この結果より，CNT はデータセットから密にサンプルされた配列を効率よく除去できていることがわかる．

3.3 リサンプリングの実行時間

ヒト H3N2 インフルエンザウイルスの HA の 4,655 本の配列から 1,000 本の配列をリサンプリングする際の, CNT, NHC, kMC および ZAS05 の実行時間(単位は秒)を表 1 に示す．CNT と ZAS05 は入力が系統樹であるため，系統樹作成のための時間が余計にかかっている．

表 1 各リサンプリングアルゴリズムの実行時間

Table 1　Execution times of the resampling algorithms.

	algorithm			
	CNT	ZAS05	NHC	kMC
Constructing a distance matrix	183	183	183	183
Constructing a tree	1,072	1,072	0	0
Resampling	54	2,011	198	1
Reconstructing a tree	16	16	16	16

4. 考察

データベースに蓄積された遺伝子情報の巨大化により，データセット全体を用いた解析に多大な時間が必要となっている．リサンプリングを行うことによりコンパクトな配列データセットを手に入れることができ，懐石に必要な計算時間も抑えることが可能となる．また，サンプリングバイアスが解析に影響を与えることも考えられる．もしデータセットに存在するサンプリングバイアスを取り除くことが出来れば，元のデータセットを用いるよりも正確な解析をすることが期待できる．以上から，リサンプリングにおいてはその実行時間よりもサンプリングバイアスを取り除く能力が重要であると考えられる．

本論文において提案した CNT が優れた性能を示した要因の 1 つとして, CNT が年代の新しい配列を多く取り除き，古い配列を残す傾向があることが挙げられる．

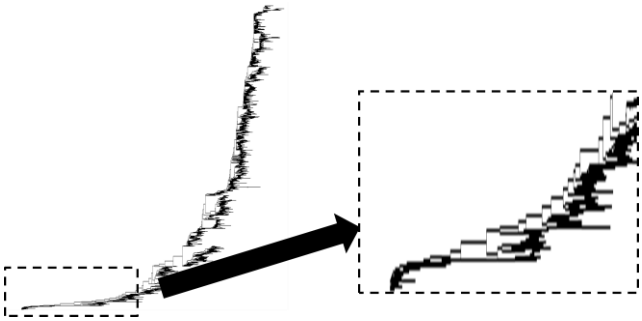

図 6　元のデータセットから作成された系統樹

Figure 6　Phylogenetic tree constructed from the original dataset.

元のデータセットから作成された系統樹(図 6)を見ると, 古い年代で分離された配列は数が少なく，しかも系統樹の根のあたりに固まって存在しており，近隣がペアを成していない．このことが原因で，CNT が新しい配列を多く選定していると考えられる．

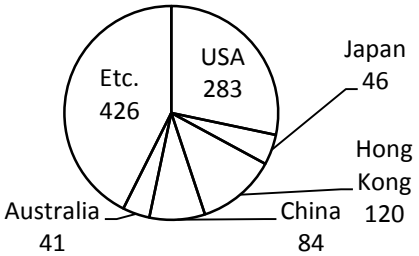

図 7　CNT によってリサンプルされた 1,000 本の配列の分離国の分布

Figure 7　Distribution of isolation countries of 1,000 sequences resampled by CNT.

CNT によってリサンプルされた 1,000 本の配列の分離された国の分布に注目すると，香港や中国から分離された配列の割合が元のデータセットよりも高くなり，逆にアメリ

カや日本からの配列の割合が減少するという興味深い現象が見られた(図 7). この現象は，アメリカや日本が他国に比べて高いサーベイランス能力を持っており，そのためサンプリングが他国よりも密に行われている可能性があることを示唆している．

5. 結論

本論文において，配列のリサンプリングを目的として提案したCNTアルゴリズムがヒトH3N2インフルエンザウイルスの HA の遺伝子配列データにおいて，配列の同一性およびサンプリングバイアスの除去能力の点で他のアルゴリズムよりも高い性能を示した．配列データは生命情報科学の研究の基礎材料として広く用いられており，サンプリングバイアス現象を目的として提案した CNT アルゴリズムの様々な配列データへ適用されることが期待できる．実際，ポリオーマウイルスやC型肝炎ウイルスなどの病原体のデータセットの他，トランスポゾンのデータにも適用が可能であり，それぞれの研究に大きな寄与ができるものと考えている．

謝辞 本研究はグローバル COE プログラム「人獣共通感染症国際共同教育研究拠点の創成」, 科学研究費若手研究(B)「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」, 感染症研究国際ネットワーク推進プログラム(J-GRID)，さきがけプロジェクト(PRESTO)および戦略的創造研究推進事業(SORST)より支援を受けている．

参考文献

1) Bao, Y. et al.: The Influenza Virus Resource at the National Center for Biotechnology Information, J. Biol., Vol.82, No.2, pp.596-601 (2008).
2) HIV sequence compendium 2010: Los Alamos National Laboratory, Los Alamos (2010).
3) Pevzner, P. A.: Multiple Alignment, In Computational Molecular Biology, The MIT Press, Cambridge, 2nd edition (2001).
4) Saitoh, N. and Nei, M.: The Neighbor Joining Method: A New Method for Reconstructing Phylogenetic Trees, Mol. Biol. Evol., Vol.4, No.4, pp.406-435 (1987).
5) Altsuhul S. F., et al.: Basic Local Alignment Search Tool, J. Mol. Biol., Vol.215, No.3, pp.403-410 (1990).
6) Zaslavsky, L. et al.: Visualization of Large Influenza Virus Sequence Datasets Using Adaptively Aggregated Trees with Sampling-Based Subscale Representation, BMC Bioinformatics, Vol.9, No.237 (2008).
7) Socal, R. and Michener, C.: A Statistical Method for Evaluating Systematic Relationships, University of Kansas Science Bulletin, Vol.38, pp.1409-1438 (1958).
8) Vinod, H.: Integer Programming and the Theory of Grouping, J. Amer. Statist. Assoc., Vol.64, pp.506-519 (1969).
9) Katoh, K., et al.: MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform, Nucl. Acid. Res., Vol.30, No.14, pp.3059-3066 (2002).
10) Felstein, J.: PHYLIP: Phylogeny Interface Package, University of Washington, Seattle (1993).
11) Jukes, T. H. and Cantor C. R.: Evolution of Protein Molecules, In Mammalian Protein Metabolism, Academic Press, New York, pp.21-132 (1969).
12) Saiki, R. H., et al.: Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase, Science, Vol.239, No.4839, pp.487-491 (1988).