G-024

木構造のランダム生成と学習
Random Generation and Inductive Learning of Tree Structures

和佐 州洋† 有村 博紀† 伊藤 公人‡
Kunihiro Wasa Hiroki Arimura Kimihito Ito

概要: 本稿では，分岐数を限定しない根付き木の確率的生成モデルを提案する．

1. はじめに

感染症研究において，あるウイルスの進化系統樹の形状は，そのウイルスの伝播の様式と密接な関係があると考えられている．そこで，与えられた木によく似た形状を持つ木構造のランダム生成手法とパラメータ推定を行う．

本稿では，新たに養分分配法を用いた GenTreeDN と中華食堂過程を用いた木生成法 GenTreeCRP の 2 つを提案する．従来，定数分岐数の根付き木の生成が主であったが，GenTreeCRP は分岐数を限定しない木の確率的生成モデルである．目視によるパラメータ探索を用いた 3 つの確率的木生成法の比較実験では，GenTreeCRP が，目標のインフルエンザウイルスの系統樹に最も似た木構造を生成した．さらに，最尤法によるモデルパラメータ推定の実験を行った．

2. 準備

2.1 根付き木

木とは，**グラフ**で表現されるデータ構造の一種である．有向グラフ $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ は有限個の**ノード**の集合 $\boldsymbol{V}$ と，ノード同士を結ぶ**有向枝**の有限集合 $\boldsymbol{E}$ を持つ．ノード列 $v_1, v_2, \ldots, v_k$ が $(v_i, v_{i+1}) \in E, i = 1, 2, \ldots, k-1$ を満たすとき，v_1 から v_k への**路**という．**根付き木**とは閉路を持たない有向グラフ $\boldsymbol{T} = (\boldsymbol{V}, \boldsymbol{E})$ で，**根**という特別なノードから，他の任意のノードへの路が存在するものをいう．木が枝 $(u, v) \in \boldsymbol{E}$ を持つとき，ノード u はノード v の**親**であるといい，v は u の**子**であるという．ノード v の子の数を $c(v)$ とし，v の k 番目の子を $v[k]$ とする．ノード u からノード v への路が存在するとき，u を v の**先祖**といい，v を u の**子孫**という．同じ親を持つノード集合を**兄弟**という．子を持たないノードを**葉**という．木 $\boldsymbol{T}$ のノード x を根とする部分木を $\boldsymbol{T}(x)$ とする．木 $\boldsymbol{T}$ に含まれるノードの総数を木の**サイズ**とし，$|\boldsymbol{T}|$ とする．根からノード u への路の長さを**深さ**といい，$d(u)$ とする．以下，特に混乱がない場合，木は根付き木を示すものとする．また，アルゴリズムで必要ならば，適宜，兄弟間に順序を仮定する．

3. バランスしない木の確率的生成アルゴリズム

以下では木の確率的生成アルゴリズムを与える．必ずしもバランスしない木を生成するために各アルゴリズムは工夫している．全てのアルゴリズムは，正整数 N と適当なモデルパラメータを受け取り，サイズ N の根付き木を生成する．確率分布に従うサンプリングが定数時間と仮定すると，どのアルゴリズムの総計算時間も，生成した木のサイズ N に対して $O(N^2)$ である．

† 北海道大学大学院情報科学研究科, Hokkaido University
‡ 北海道大学人獣共通感染症リサーチセンター, Hokkaido University

3.1 従来手法: 木 DLA

拡散律速凝集 (Diffusion Limited Aggregation) は，Witten と Sander(1981) によって提案された凝集の物理学モデルである．DLA では，無限格子 $\mathbb{Z}^d$ 上に置かれた結晶に対して，新しい粒子を無限遠点からランダムウォークさせ，それが付着することで結晶が成長する．Martin 他 [3] は，これを無限 d 分木上に拡張した木 DLA を提案した．この手法は分岐数 d の木を生成する．彼らは，木 DLA が，パラメータ $\alpha > 0$ に対して，木の分岐数 d 未満のとき各ノード v に確率 $p(v) = \beta\alpha^{-depth(v)}$ で新たな子ノードを追加して，木を成長させる Algorithm 1 の生成手法 (ここでは GenTreeDLA) に等価なことを示した．ただし，β は正規化定数である．

Algorithm 1 木 DLA を用いた木生成法

```
1: procedure GenTreeDLA(N, α)
2:     T ← tree with the root node only;
3:     for i ← 1 ... N − 1 do AddTreeDLA(T, α);
4:     return T;
5: procedure AddTreeDLA(T, α)
6:     select v ∈ T with probability p(v) = βα^{−depth(v)};
7:     add child to v;
```

3.2 提案手法 1: 養分分配法

次に，養分分配を用いた木生成法 GenTreeDN を (Algorithm 2) を提案する．この手法は分岐数 d の木を生成する．再帰手続き AddTreeDN は，d 次元ディリクレ分布の任意のパラメータ $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_d)$ に対して，各ノード v とその部分木サイズを受け取り，その部分木サイズの組 $(N_1, \ldots, N_d)$ を d 次元ディリクレ分布 $Dir(\boldsymbol{\alpha})$ に従って生成する．これを再帰的に繰り返し，木を生成する．

Algorithm 2 養分分配を用いた木生成法 GenTreeDN

```
1: procedure GenTreeDN(N, d, α)
2:     T ← tree with the root node only;
3:     AddTreeDN(root(T), N, d α);
4:     return T;
5: procedure AddTreeDN(v, n, d, α)
6:     select a d-vector (N_1, ..., N_d) such that Σ_i N_i =
       n − 1 according to the distribution Dir(α);
7:     for i ← 1 ... d do AddTreeDN(v[i], N_i, d, α);
```

3.3 提案手法 2: 中華食堂過程を用いた木生成法

中華食堂過程 (Chinese Restaurant Process, CRP) は，個体のグループの族を生成する確率過程である [1]．最後に提案する木生成法 GenTreeCRP(Algorithm 3) は，この中華食堂過程を階層的に用いて，分岐数が限定されない木をトップダウンに生成する．パラメータ $\gamma > 0$ に対して，再帰手続き AddTreeCRP は，根からスタートし，木を下降しながら，次のように新しいノードを追加する．各頂点 v において，AddTreeCRP は，確率 $p_{add}(v) = \frac{\gamma}{\gamma+|v|-1}$ で v に新たな子ノードを追加し，確率 $p_{visit}(v) = \frac{|v[i]|}{\gamma+|v|-1}$ でその子 $v[i]$ を訪問し，再帰的に計算を行う．根ノードのみ

Algorithm 3 CRP を用いた木生成法 GenTreeCRP

procedure GenTreeCRP(N, γ)
  $T \leftarrow$ tree with the root node only;
  for $i \leftarrow 1 \ldots N-1$ **do** AddTreeCRP(root($\boldsymbol{T}$), γ);
  return T;
procedure AddTreeCRP(v, γ)
  $k \leftarrow c(v)$;
  for $j \leftarrow 1 \ldots k$ **do** $N_j \leftarrow |T(v[j])|$;
  $N \leftarrow N_1 + \cdots + N_k$
  select $1 \le i \le k+1$ by $\begin{cases} \frac{\gamma}{N+\gamma}, i = k+1 \\ \frac{N_i}{N+\gamma}, i = 1, 2, \ldots, k \end{cases}$
  if $i = k+1$ **do** add to v a new child $v[k+1]$ as the youngest sibling ;
  else do AddTreeCRP($v[i], \gamma$);

の木に対して，この追加操作を $N-1$ 回行うことでサイズ N の木を生成する．

4. 実験

4.1 データと方法

データは NCBI Influenza Virus Resource から H1N1 型インフルエンザ 4240 株を取得し近隣結合法で作成した系統樹を入力とした．プログラムの実装には C++ と数値解析ソフト GNU Octave を使用した．また，木構造の描画には Dendroscope[4] を使用した．

4.2 実験 1: アルゴルズムの比較

目視によるパラメータ推定により，各アルゴリズムによってインフルエンザウイルス株の系統樹に似た木構造を生成し，生成された木構造を比較した．なお，使用したパラメータの値は，予備実験で目視によるパラメータ探索で求めた．図 2,3,4 に結果の木構造を示す．図から，GenTreeCRP($\gamma = 0.1$) が最もよく似た木構造を生成したことがわかる．

4.3 実験 2:GenTreeCRP のパラメータ推定

インフルエンザウイルス株の系統樹から最尤法によって，パラメータ推定を行った．最尤パラメータは $\gamma = 0.45$ であった．図 5 に結果の木構造を示す．

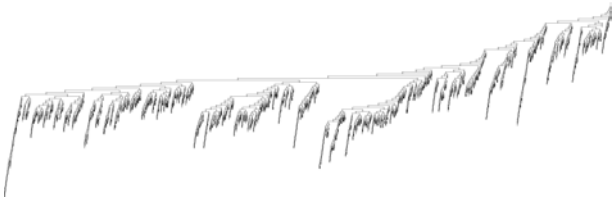

図 1　H1N1 インフルエンザウイルス株の進化系統樹

5. まとめ

本稿では，木構造のランダム生成とパラメータ推定を考察した．今後，このような手法は，様々なウイルス種の系統樹の特徴を調べたり，相互比較するのに役立つと期待される．

参考文献

[1] Aldous, D. J. Exchangeability and related topics. In *École d'Été de probabilités de Saint-Flour, XIII– 1983*, pages 1-198. Springer, Berlin, 1985.

[2] Aldous, D. J. *Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today.* Statistical Science, 16(1), 23-34, 2001

[3] Martin T. Barlow, Robin Pemantle, Edwin A. Perkins, *Diffusion-limited aggregation on a tree*, Probability Theory and Related Fields, 107:1-60, 1997.

[4] Daniel H Huson, Daniel C Richter, Christian Rausch, Tobias Dezulian, Markus Franz and Regula Rupp. *Dendroscope: An interactive viewer for large phylogenetic trees*, BMC Bioinformatics. 2007 Nov 22;8(1):460

[5] 柳橋史成, 伊藤公人, 有村博紀, 木の最適ラベリング問題とその進化系統樹への応用. 情報処理学会研究報告, BIO, バイオ情報学, 2009(25), 29-32, 一般社団法人情報処理学会. 2009.

図 2　GenTreeDLA, $\alpha = 0.8, N = 4240$

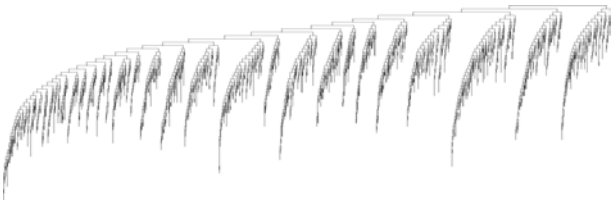

図 3　GenTreeDN, $\boldsymbol{\alpha} = (100, 10), N = 4240$

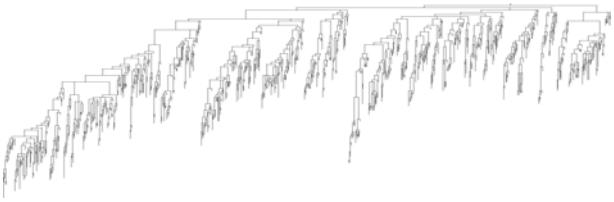

図 4　GenTreeCRP, $\gamma = 0.1, n = 4240$

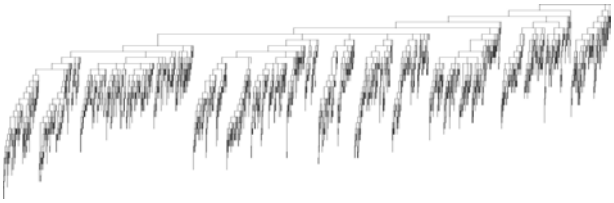

図 5　GenTreeCRP, $\gamma = 0.45, n = 4240$