

環境中の DNA 配列の集団のダイナミクスの解析

小谷野 仁^{1,a)} 澤田 和典² 山本 希¹ 山田 拓司¹

概要: あるアルファベット $A = \{a_1, \dots, a_z\}$ から作られる文字列の全体 A^* は, 数学の立場から見ると, Levenshtein 距離などの編集距離によって距離空間に, 接続によって半群になり, 非可換な位相半群をなす. また, 1 つの環境中の生物群集が持つ DNA やある遺伝子の塩基配列の全体 (母集団) の分布は, アルファベット $A = \{a, c, g, t\}$ に対する位相半群 A^* 上の確率関数として表される. 本稿では, 私達の以前の 2 つの研究: (i) 環境中の DNA 配列の集団にその環境から掛かる淘汰圧の分布をその混合モデルによって表すように設計して, A^* 上に導入した Laplace 様分布という確率分布とその混合モデルに対する統計的推定論を展開した研究と, (ii) 1 つの環境中の生物群集が持つ DNA 配列の集団が, 配列中に確率的に起こる突然変異と, 周囲の環境から掛かる淘汰圧の下で時間発展していく様子を記述する偏微分方程式を導出し, その数理解析を行った研究, を基礎として取り組んだ, 環境中の微生物群集が持つ 16S リボソーム RNA 遺伝子の集団のダイナミクスを予測した研究の結果を述べる. また, 遺伝子の母集団のダイナミクスに対して変化速度や方向持続性などの特徴量を導入し, それらを遺伝子の標本から統計的に推定する方法を開発して, 遺伝子の集団のダイナミクスの特徴を調べた研究の結果を述べる.

Analyzing the Dynamics of a Population of DNA Sequences in an Environment

1. はじめに

1 つの環境中の生物群集は, 群集中の他の個体や周囲の環境と相互作用しながら時間発展していく. 本稿では, 私達はこのダイナミクスに興味がある. アルファベット $A = \{a, c, g, t\}$ 上の文字列がなす非可換位相半群を A^* によって表す. そうすると, DNA や遺伝子は A^* の元であり, 1 つの環境中の生物群集が持つ DNA やある遺伝子の全体の分布は, A^* 上の確率関数として表される. そこで, 本稿では, 私達は, 1 つの環境中の生物群集を, それが持つ DNA の集団や 16S リボソーム RNA 遺伝子の集団として捉えることにより, そのダイナミクスを予測したり, ダイナミクスの特徴を調べる問題にアプローチする.

2. 基礎になっているこれまでの研究

本節では, 本稿で述べる研究が基づいている, 私達の以前の 2 つの研究について簡単に述べる.

2.1 環境から DNA の集団への淘汰圧を推定する

A^* 上には, これまで, 実数の集合 \mathbb{R} 上の正規分布や自然数の集合 \mathbb{N} 上の 2 項分布のようなパラメトリックな分布は導入されていなかった. Koyano *et al.* [1] は, 環境中の DNA の集団にその環境から掛かる淘汰圧の分布をその混合モデルによって表すように設計された Laplace 様分布という確率分布を A^* 上に導入し, その基本的な性質を調べることから始めて, その混合モデルのパラメーターを推定する方法を開発し, それに対して数理的基礎付けを与え, 数値実験と実データ解析を行って, その有効性を確かめた.

2.2 DNA の集団の時間発展を偏微分方程式で表す

Koyano *et al.* [2] は, 拡散方程式をモチーフにして, A^* 上で, 1 つの環境中の生物群集が持つ DNA の集団が, 配列中に確率的に起こる突然変異と, 周囲の環境から掛かる淘汰圧の下で時間発展していく様子, すなわち DNA の集団の進化を記述する偏微分方程式を導出した. また, モデルの数理解析を行って, 集団が分化して新しい種が作られるための条件や, 集団が平衡状態を維持し長期間に渡って変化しないための条件を示した.

¹ 東京工業大学
² 株式会社ぐるなび
^{a)} koyano.h.aa@m.titech.ac.jp

3. DNA の集団のダイナミクスの解析

第 2.1 節で述べた方法と第 2.2 節で述べた理論を組み合わせ、環境中の微生物群集が持つ 16S リボソーム RNA 遺伝子の集団のダイナミクスの予測を試みる。図 1 は、植物 *Brassica rapa* subvar. *hiroshimana* Kitam. の周辺環境中の微生物群集から収集した 16S リボソーム RNA 遺伝子の集団を可視化したものである。各遺伝子はアルファベット $A = \{a, c, g, t\}$ 上の文字列の位相半群 A^* の元であるから、そのままでは図示できない。ここでは、Levenshtein 距離で遺伝子の間の差異を定量し、多次元尺度法を用いて標本中の各遺伝子を平面 \mathbb{R}^2 上に配置した後、 \mathbb{R}^2 を長方形の区画に区切り、各区画に属する遺伝子の数を数えてヒストグラムを作成することにより可視化した。図 2 は、図 1 の環境を人工的に高塩濃度の環境に改変した後、その環境中の微生物群集から収集した 16S リボソーム RNA 遺伝子の集団を同様の方法で可視化したものである。同一の遺伝子は同一の座標となって比較できるように、これらの図は作成してある。環境の改変により、微生物群集の構造が著しく変化したことが分かる。図 3 は、図 2 の遺伝子の集団に掛かる淘汰圧を、第 2.1 節で述べた方法を用いて推定し、可視化したものである。淘汰圧は A^* 上の非負の実数値関数であるから、これを \mathbb{R}^2 上に射影して図示してある。図 4 は、第 2.2 節で述べた理論モデルから、図 2 の遺伝子の集団に掛かる推定された淘汰圧 (図 3 参照) の下での図 1 の遺伝子の集団の時間発展を計算した結果を、上記の方法で可視化して示したものである。図 1 の遺伝子の集団が図 2 の集団へと徐々に変化していくことが見てとれる。

植物 *Solanum melongena kouki* の周辺環境を人工的に高塩環境に改変する前の 1 時点と改変した後の 1, 7, 15, 30 日の 4 時点で 16S リボソーム RNA 遺伝子の標本を収集して、同様の数値実験を行った場合でも、理論から得られる遺伝子の集団が、観測された遺伝子の集団が示すダイナミクスと同じパターンで時間発展することを示された。更に、私達は、遺伝子の母集団のダイナミクスに対して変化速度や方向持続性などの特徴量を導入し、それらを遺伝子の標本から統計的に推定する方法を開発して、上記の 2 つの数値実験における遺伝子の集団のダイナミクスの特徴を調べた。発表では、これらの結果も報告する。

参考文献

- [1] Koyano, H., Hayashida, M. and Akutsu, T.: Optimal string clustering based on a Laplace-like mixture and EM algorithm on a set of strings, *Journal of Computer and System Sciences* (2019).
- [2] Koyano, H. and Yano, K.: Evolutionary model of a population of DNA sequences through the interaction with an environment and its application to speciation analysis, *arXiv:1706.01182[q-bio.PE]*.

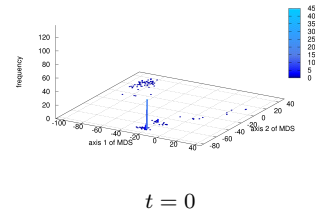


図 1 *Brassica rapa* subvar. *hiroshimana* Kitam. の周辺環境中の微生物群集から収集した 16S リボソーム RNA 遺伝子の集団のヒストグラム。

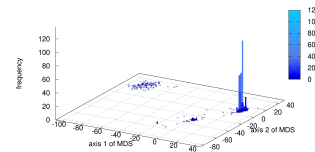


図 2 図 1 の環境を人工的に高塩環境に改変した後に微生物群集から収集した 16S リボソーム RNA 遺伝子の集団のヒストグラム。

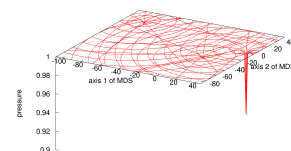


図 3 図 2 に示された遺伝子の集団に掛かる推定された淘汰圧。

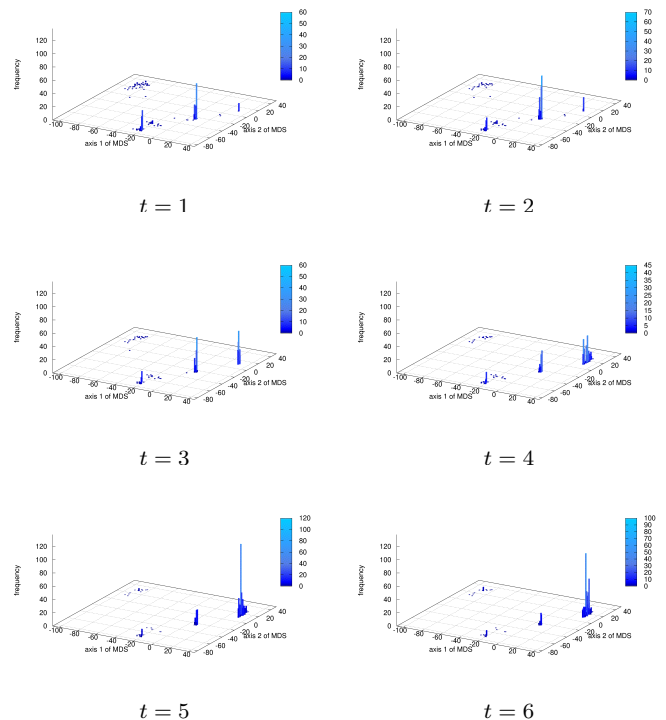


図 4 図 1 の集団を初期状態として、図 2 の集団への推定された淘汰圧 (図 3 参照) の下で理論モデルから計算される配列の集団の時間発展。徐々に図 2 の集団に近づいていく。