7F-04

DNA 配列の集団の時間発展を記述する偏微分方程式とその種分化の分析への応用 *

小谷野 仁 †
東京工業大学 生命理工学院

アルファベット $A = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$ 上の文字列の集合を A^* によって表す. A^* は連接 $\cdot$ (A^* の 2 つの元を繋げる操作) によって半群, Levenshtein 距離 d_L によって距離空間であり, 非可換な位相半群をなす. 本研究において, 我々は, この空間 A^* 上で, 微生物群集のように周囲の環境と相互作用する集団が持つ DNA 配列の集団の時間発展のモデルを構築し, それを応用して, 種分化の分析に取り組んだ.

時刻 $t \in [0, \infty)$ におけるある環境中の DNA 配列の集団を $S(t)$ によって表す. $s \in S(t)$ ならば, $s \in A^*$ である. X が集合である時, $|X|$ は X の元の数を表し, $s \in A^*$ の時, $|s|$ は s の長さ (すなわち, s を構成する A の元の数) を表すとする. $n(t) = |S(t)|$ とおき, $S(t) = \{s_1(t), \cdots, s_{n(t)}(t)\}$ と書く. $R(t)$ を相等 $=$ に関する $S(t)$ の商集合 $S(t)/=$ の完全代表系 (すなわち, $S(t)$ の配列のリスト) とする. $s \in R(t)$ の同値類の元の数 (すなわち, $s \in R(t)$ と等しい $S(t)$ の配列の数) を $x(s,t)$ によって表し, $q(s,t) = x(s,t)/n(t)$ と定める. $Q(t) = \{q(s,t) : s \in A^*\}$ とおき, $Q(t)$ を $S(t)$ の相対頻度分布と呼ぶ. 当面, 集団の大きさ $n(t)$ は外生的に与えられているとする.

$\mathcal{E}$ を地球上の様々な地理的領域に存在し得る環境の全体とする. ある配列 $s \in A^*$ がある環境 $E \in \mathcal{E}$ の下に存在する場合に, s が E から受ける淘汰圧を $p(s,E)$ によって表す. 従って, 淘汰圧は, 環境 $E \in \mathcal{E}$ が与えられると定まる関数 $p(\,\cdot\,, E) : A^* \to [0, \infty)$ である. 以下で定式化されるように, $p(s,E)$ が大きい程, s の E への適応度は低く, その結果, s の子の数は少ない. 逆の場合, 逆である. どの $s \in A^*$ もその淘汰圧の下では E の中で生存できない淘汰圧の水準があるとし, その水準を p_E によって表す. $p(s,E)$ のモデルについてはここでは略し, 発表において述べる.

定数 $\gamma \in (0,1]$ に対して, 単位時間 $[t, t+\Delta t]$ の間に, $S(t)$ の中の γ の割合の配列が子を残して死ぬとする. この設定では, $\gamma < 1$ ならば, 配列により寿命は異なるが, 世代重複までは考慮に入れられていない. $[t, t+\Delta t]$ の間に子を残して死ぬ $S(t)$ の配列の集合を $\hat{S}(t)$ によって表し, $\hat{S}(t)$ は次の規則 C_1 に従って定まるとする. C_1: 略 (発表において述べる). g_t を, $S(t)$ が与えられた時, 規則 C_1 によって定まる $\hat{S}(t) \subset S(t)$ を返す写像とする. すなわち, $\hat{S}(t) = g_t(S(t))$.

$\hat{R}(t)$ を相等 $=$ に関する $\hat{S}(t)$ の商集合 $\hat{S}(t)/=$ の完全代表系 (すなわち, $\hat{S}(t)$ の配列のリスト) とする. $s \in A^*$ と等しい $\hat{S}(t)$ の配列の数を $\hat{x}(s,t)$ によって表す. 関数 $f(\,\cdot\,, t, E) : \hat{R}(t) \to [0,1]$ を, $p(s,E) < p_E$ の時,

$$f(s,t,E) = \frac{p_E - p(s,E)}{\sum_{s' \in \hat{R}(t)} (p_E - p(s',E))},$$

そうでない時, $f(s,t,E) = 0$ と定義する. 任意の $s \in \hat{R}(t)$ に対して $0 \le f(s,t,E) \le 1$ であって, $\sum_{s \in \hat{R}(t)} f(s,t,E) = 1$ である.

各 $s \in \hat{R}(t)$ に対して, s と等しい $\hat{S}(t)$ の配列のうちの 1 つの子の数 $o(s,t)$ は, 次の規則 C_2 によって定まるとする. C_2: 略 (発表において述べる). $\hat{S}(t)$ の配列が残す子配列の数を $\tilde{n}(t)$, それらの集合を $\tilde{S}(t)$ によって表す. 従って, $\tilde{n}(t) = |\tilde{S}(t)|$. $f(s,t,E)/\hat{x}(s,t)$ と $o(s,t)$ は, それぞれ $s \in \hat{R}(t)$ と等しい $\hat{S}(t)$ の配列のうちの 1 つの相対的適応度と適応度と見なすことが出来る.

m を, A^* の 1 つの元を複数回与えた時, それを常に A^* の同一の元に対応させるとは限らない, A^* 上のランダムな変換とする. $s, s' \in A^*$ に対して, 入力として s が与えられた時に, m が s' を出力する確率を $\mu(s,s')$ によって表す. $s, s', s'' \in A^*$ に対して, m が次の 3 つの条件を満たす時, m を突然変異作用素と呼ぶ: (i) $d_L(s,s') < d_L(s,s'')$ なら

*Evolutionary model of a population of DNA sequences interacting with a surrounding environment and its application to speciation analysis

†Hitoshi Koyano, School of Life Science and Technology, Tokyo Institute of Technology

ば, $\mu(s,s') > \mu(s,s'')$, (ii) $d_L(s,s') = d_L(s,s'')$ ならば, $\mu(s,s') = \mu(s,s'')$, 及び (iii) m によって生成される突然変異を含む配列は独立である. ここでは略すが, 発表においては, A^* 上の連接 $\cdot$ を用いて, Levenshtein 距離 d_L に対してこれらの条件を満たす m を具体的に構成する. 時刻 t における突然変異作用素を, $\tilde{n}(t)$ 個の文字列の集合 $\{s_1, \cdots, s_{\tilde{n}(t)}\}$ に, 次のように成分毎に作用する $\tilde{n}(t)$ 個の突然変異作用素の集合 $m_t = (m_t^{(1)}, \cdots, m_t^{(\tilde{n}(t))})$ として定義する:

$$m_t(s_1, \cdots, s_{\tilde{n}(t)}) = (m_t^{(1)}(s_1), \cdots, m_t^{(\tilde{n}(t))}(s_{\tilde{n}(t)})).$$

子配列が Levenshtein 距離 1 に相当する突然変異を含む確率を $\pi \in (0,1)$ によって表す.

任意の $n \in \mathbb{N}$ ($\mathbb{N}$ は 0 を含む自然数の集合を表す) と $s \in A^*$ に対して, 記号 $\bigcup^n\{s\}$ ($\bigcup$ が上付きの添え字のみを持つ) を, $n = 0$ の時 $\bigcup^n\{s\} = \emptyset$ と定め, $n \geq 1$ の時 $\bigcup^n\{s\} = n$ 個の s の集合と定める. そうして, 時刻 t における複製作用素 $r_t = (r_t^{(1)}, \cdots, r_t^{(\hat{n}(t))}) : (A^*)^{\hat{n}(t)} \to (A^*)^{\tilde{n}(t)}$ を

$$\begin{aligned} r_t(s_1, \cdots, s_{\hat{n}(t)}) &= (r_t^{(1)}(s_1), \cdots, r_t^{(\hat{n}(t))}(s_{\hat{n}(t)})) \\ &= \bigcup_{i=1}^{\hat{n}(t)} \bigcup^{o(s_i,t)} \{s_i\} \end{aligned}$$

と定義する. 上の式の最初の等号は, 複製作用素は成分毎に作用することを, 2 番目の等号は, 複製作用素の各成分は受け取った配列をその適応度に等しい数だけ複製することを示している.

最後に, 時刻 t における生成作用素 G_t を $G_t(S(t)) = S(t) \setminus g_t(S(t)) \cup m_t \circ r_t \circ g_t(S(t))$ と定義する. ここで, $\circ$ は合成写像を表す. このように定義すると, 集団 $S(t+\Delta t)$ は集団 $S(t)$ から次の仕方で生成される:

$$\begin{aligned} &m_t \circ r_t \circ g_t(S(t)) = m_t \circ r_t(\hat{S}(t)) \\ &= m_t\big(r_t^{(1)}(\hat{s}_1(t)), \cdots, r_t^{(\hat{n}(t))}(\hat{s}_{\hat{n}(t)}(t))\big) \\ &= m_t\left(\bigcup_{i=1}^{\hat{n}(t)} \bigcup^{o(\hat{s}_i(t),t)} \{\hat{s}_i(t)\}\right) \\ &= \big\{m_t^{(1)}(s'_1), \cdots, m_t^{(\tilde{n}(t))}(s'_{\tilde{n}(t)})\big\} \\ &= \{\tilde{s}_1(t), \cdots, \tilde{s}_{\tilde{n}(t)}(t)\} = \tilde{S}(t), \\ &G_t(S(t)) = S(t) \setminus g_t(S(t)) \cup m_t \circ r_t \circ g_t(S(t)) \\ &= S(t) \setminus \hat{S}(t) \cup \tilde{S}(t) = S(t+\Delta t). \end{aligned}$$

ここで, $\bigcup_{i=1}^{\hat{n}(t)} \bigcup^{o(\hat{s}_i(t),t)} \{\hat{s}_i(t)\} = \{s'_1, \cdots, s'_{\tilde{n}(t)}\}$ であって, 各 $i = 1, \cdots, \tilde{n}(t)$ に対して $m_t^{(i)}(s'_i) = \tilde{s}_i(t)$ である.

確率的な突然変異により DNA 或いは遺伝子配列の 2 つの塩基 (A の 2 つの文字) の間に挿入され得る配列の長さには上限があると仮定し, それを c によって表す. $s \in A^*$ に対して $\ell(s) = c(|s|+1)+|s|$ と定める. $\ell(s)$ は s の子配列に起こり得る突然変異の回数の上限である. $d \in \mathbb{N}$ に対して $V(s,d) = \{s' \in A^* : d_L(s,s') = d\}$ とおく. ${}_nC_r$ は n 個のものから r 個のものを取る組合せの数を表す. $\Delta t \to 0$ の時, 大雑把には, $n(t+\Delta t) \to n(t)$ と $\hat{x}(s,t) \to 0$ から, $q(s,t) - x(s,t)/n(t+\Delta t) \to 0$ と $\hat{x}(s,t)/n(t+\Delta t) \to 0$ となる. 上述の設定の下で, 次の結果が得られる.

任意の $t \in [0,\infty)$ と $s \in A^*$ に対して, $\Delta t \to 0$ の時の

$$\frac{1}{\Delta t}\frac{\hat{x}(s,t)o(s,t)}{n(t+\Delta t)},\ \frac{1}{\Delta t}\left(q(s,t) + \frac{\hat{x}(s,t) - x(s,t)}{n(t+\Delta t)}\right)$$

の極限 $b(s,t)$ と $c(s,t)$ が存在するならば, $S(t)$ の相対頻度分布 $q(s,t)$ の時間発展は, 偏微分方程式

$$\begin{aligned} \frac{\partial q(s,t)}{\partial t} &= -c(s,t) + b(s,t)(1-\pi)^{\ell(s)} \\ &+ \sum_{1 \leq d < \infty} \sum_{s' \in V(s,d)} b(s',t)\frac{\ell(s')C_d\pi^d(1-\pi)^{\ell(s')-d}}{|V(s',d)|} \end{aligned}$$

によって記述される. $b(s,t)$ は, 時刻 t において子を残して死ぬ, s と等しい配列の子配列の $S(t)$ における相対頻度, $c(s,t)$ は, 時刻 t において子を残して死ぬ, s と等しい配列の相対頻度を表す.

発表においては, 上述のモデルの数理解析とそれに基づいた数値実験を行って, 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でいるための条件を調べた結果を述べる. この発表は [1] と [2] に基づいている.

引用文献

[1] H. Koyano, M. Hayashida, and T. Akutsu. Optimal string clustering based on a Laplace-like mixture and EM algorithm on a set of strings. arXiv:1411.6471[math.ST].

[2] H. Koyano and K. Yano. Evolutionary model of a population of DNA sequences interacting with a surrounding environment and its application to speciation analysis. arXiv:1706.01182[q-bio.PE].