Vol.2021-MPS-132 No.23
2021/3/2

Laplace 様混合モデルの基準化定数の計算

小谷野 仁[1,a)]　林田 守広[2]

概要：n 次元 Euclid 空間 $\mathbb{R}^n$ における超球と超球面の体積は, よく知られた公式を使って簡単に計算できる. 通常の L^2 ノルムの下でのみならず, 一般の L^p ノルムの下でのこれらの体積の公式も知られている. アルファベット $A = \{a_1, \cdots, a_k\}$ から作られる文字列の集合 A^* 上にも, 拡張 Hamming 距離, 最小共通部分列距離, Levenshtein 距離, Damerau–Levenshtein 距離などの編集距離が存在するから, $\mathbb{R}^n$ における超球や超球面の対応物を定義することができる. しかし, これらの体積の公式は知られておらず, 現在のところ, これらの体積は網羅探索によって求めるしかない. A^* における球は 1 つの正則言語をなすため, その大きさは, 離散数学のみならず, 形式言語理論においても研究されてきたが, その大きさを計算する明示的な公式を得ることではなく, その大きさの成長速度を評価することが主要な目的とされてきた. このような状況の下で, 筆者等は, 1 つの環境中の微生物群集が持つ DNA や 16S rRNA 遺伝子配列の集団の時間発展を記述する偏微分方程式を導出して解析した研究 (Koyano and Yano, arXiv:1706.01182[q-bio.PE]) と, 環境からそこに生息する微生物群集が持つ DNA や 16S rRNA 遺伝子配列の集団に掛かる淘汰圧のモデル (Laplace 様混合モデルと言う) とそのパラメーターの推定方法を提案した研究 (Koyano, Hayashida, and Akutsu, 2019, *Journal of Computer and System Sciences*) の結果を組み合わせて, 数値実験において, いくつかの環境中の微生物の 16S rRNA 遺伝子配列の集団の観測された時間発展をコンピューターの中で再現できることを示した研究 (Koyano, Sawada, Yamamoto, and Yamada, submitted) において, アルファベットが $A = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ である場合に, 文字列の球面の大きさを具体的に計算する必要性に直面した. それは, Laplace 様混合モデルの基準化定数に, 文字列球面の大きさが含まれるためである. 本発表では, 一般のアルファベット $A = \{a_1, \cdots, a_k\}$ の下で A^* 上に拡張 Hamming 距離と Levenshtein 距離が定義されている場合の文字列球の大きさとその成長速度に関する私達の研究の結果を報告する.

Computing the normalizing constant of the Laplace-like mixture model

1. 問題の背景

始めに, 本研究の背景となっている, これまでの私達の研究について述べる. アルファベット $A = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ から作られる文字列 (A の元の有限列) 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) によって半群に, 拡張 Hamming 距離 ($d_{H'}$ によって表す), 最小共通部分列距離, Levenshtein 距離 (d_L によって表す), Damerau–Levenshtein 距離などの編集距離によって距離空間になり, 非可換な位相半群をなす. どの DNA 配列も A^* の元である.

1 つの環境中の微生物群集について考える. 考察する問題により微生物群集の捉え方は色々あり得るが, ここでは, それを, それが持つ DNA 配列の集団として捉える. 1 つの環境中に同一の DNA 配列を持つ微生物が複数存在し得るから, 微生物群集が持つ DNA 配列の集団を, それを構成する DNA 配列とその相対頻度の全体として捉える. この情報は, A^* 上の 1 つの確率関数が持つ情報と同じである. そこで, 1 つの環境中の微生物群集を A^* 上の 1 つの確率関数として捉えることにより, 微生物群集のダイナミクスを解析することを考える.

[9] は, 拡散方程式をモチーフにして, 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, A^* 上で定義された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でいるための条件を

[1] 農研機構農業情報研究センター
[2] 松江工業高等専門学校電気工学科
[a)] koyanoh317@affrc.go.jp

示した. 記号等の詳細は略すが, 導出されたのは次の方程式である. 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, 時刻 t における DNA 配列の集団 $S(t)$ の相対頻度分布 $q(s,t)$ の時間発展は, 偏微分方程式

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

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と解釈される.

また, [5] は, 環境から集団中の各 DNA 配列に掛かる淘汰圧を, 配列の環境標本に基づいて推定するための次のモデルを提案した. A^* 上の距離関数の集合を D によって表す. 任意の $\lambda \in A^*$, $\rho \in (0,\infty)$, 及び $d \in D$ に対して, 関数 $q_d(\,\cdot\,;\lambda,\rho) : A^* \to [0,1]$ を

$$q_d(s;\lambda,\rho) = \frac{1}{(\rho+1)\big|V_d(\lambda,d(s,\lambda))\big|}\left(\frac{\rho}{\rho+1}\right)^{d(s,\lambda)} \quad (1)$$

によって定め, 集合関数 $Q_d(\,\cdot\,;\lambda,\rho) : 2^{A^*} \to [0,1]$ を $Q_d(E;\lambda,\rho) = \sum_{s\in E} q_d(s;\lambda,\rho)$ によって定義する. ここで, $s \in A^*$, $r \in \mathbb{N}$, 及び $d \in D$ に対して

$$V_d(s,r) = \{s' \in A^* : d(s,s') = r\}$$

であって, $|X|$ は有限集合 X の元の数を表す. そうすると, $Q_d(\,\cdot\,;\lambda,\rho)$ は可測空間 $(A^*, 2^{A^*})$ 上の確率測度になることが確かめられる. $Q_d(\,\cdot\,;\lambda,\rho)$ は, 実数の集合 $\mathbb{R}$ 上の Laplace 分布と類似の性質を持つことが示される. そこで, $Q_d(\,\cdot\,;\lambda,\rho)$ を A^* 上の Laplace 様分布と名付ける. そうして, 1 つの環境中の微生物群集が持つ DNA 配列の母集団の分布を Laplace 様分布の混合モデル

$$\boldsymbol{q}_d(s;\boldsymbol{\theta}) = \sum_{g=1}^{\ell}\pi_g q_d(s;\lambda_g,\rho_g)$$

として, 配列 s への淘汰圧を $\phi(s) = 1-\boldsymbol{q}_d(s;\boldsymbol{\theta})$ としてモデル化する. ここで, $\boldsymbol{\theta} = (\lambda_1,\cdots,\lambda_\ell,\rho_1,\cdots,\rho_\ell,\pi_1,\cdots,\pi_\ell)$ である. 配列の環境標本に基づいた Laplace 様混合モデルのパラメーターの推定アルゴリズムは, [5] において述べられている. [5] では, [4], [6] の結果を用いて, その推定アルゴリズムがある正則条件の下で強一致性を持つことも示されている.

更に, [7] では, 植物 *Solanum melongene* の周辺環境と *Perilla frutescens var. crispa f. purpurea* の周辺環境を人工的に高塩環境に改変する前後に, その環境中の微生物群集が持つ 16S rRNA 遺伝子配列の集団の時間発展を観測し, これらの観測された配列の集団の時間発展が, 上記の方程式と Lapalce 様混合モデルを組み合わせて行った数値実験において再現されることを示した. [7] では, [8] の結果を使って, A^* 上の確率関数の集合 $\mathcal{P}$ 上に距離を導入し, DNA 配列の集団の時間発展に対して, 変化速度や方向持続性の概念を導入し, その計算方法を開発して, 上記の 2 つの環境中の微生物群集の時間発展の解析も行われている.

2. 考察する問題

式 (1) によって定義される Laplace 様分布の確率関数をもう一度見てみよう. この確率関数の基準化定数に含まれる $V_d(s,r)$ は $\mathbb{R}^n$ における超球面

$$S^{n-1}(\boldsymbol{x},r) = \{\boldsymbol{x}' \in \mathbb{R}^n : ||\boldsymbol{x}-\boldsymbol{x}'|| = r\}$$

の対応物であり, $|V_d(s,r)|$ は $S^{n-1}(\boldsymbol{x},r)$ の体積の対応物である. ここで, $||\,\cdot\,||$ は L^2 ノルムを表す. $\mathbb{R}^n$ における超球

$$B^n(\boldsymbol{x},r) = \{\boldsymbol{x}' \in \mathbb{R}^n : ||\boldsymbol{x}-\boldsymbol{x}'|| \le r\}$$

の体積はよく知られた公式

$$\frac{\pi^{n/2}}{\Gamma(n/2+1)}r^n$$

を用いて簡単に計算することができ, この公式から $S^{n-1}(\boldsymbol{x},r)$ の体積の公式も得られる. ここで, Γ はガンマ関数を表す. しかし, 文字列の球や球面の体積の公式は知られていない. 現在のところ, これらの体積は網羅探索によって求めるしかないが, 多くの場合, それは非常に長い時間を要する. 以下では, Laplace 様混合モデルを用いて数値実験を行うことを当初の目的として始めた, 文字列の球と球面の大きさを計算する研究の結果を述べる.

3. 主要な結果

アルファベットの大きさ $k \in \mathbb{Z}^+$ ($\mathbb{Z}^+$ は正の整数の集合) を明示して, 文字列球面を $V_{k,d}(s,r)$ によって表す. また, アルファベットの大きさが k である時の文字列球を $U_{k,d}(s,r)$ によって表す. すなわち,

$$U_{k,d}(s,r) = \{s' \in A^* : d(s,s') \le r\}.$$

$s \in A^*$ に対して, $|s|$ は s の長さを表す.

まず, A^* 上に拡張 Hamming 距離 $d_{H'}$ が定義されている場合の文字列球の体積に関する結果を述べる. 任意の $k \in \mathbb{Z}^+, s \in A^*$, 及び $r \in \mathbb{N}$ ($\mathbb{N}$ は自然数の集合) に対して, $|U_{k,d_{H'}}(s,r)|$ は式 (2) によって与えられる. 従って, $|U_{k,d_{H'}}(s,r)|$ は, (i) $|s|$ と r が固定されている時, k に関して $\Theta(poly(k))$ に属し, (ii) k と r が固定されている時, $|s|$

$$|U_{k,d_{H'}}(s,r)| = \begin{cases} 1 & r=0 \text{ の時,} \\ 1+\sum_{r'=1}^{r}\left\{k^{r'}\sum_{r_1=0}^{r'-1}C(|s|,r_1)\left(\frac{k-1}{k}\right)^{r_1} \right. & \\ \qquad \left. +\sum_{r_1=0}^{r'}C(|s|-r'+r_1,r_1)(k-1)^{r_1}\right\} & 1\le r\le |s| \text{ の時,} \\ 1+\left(2-\frac{1}{k}\right)^{|s|}\frac{k(k^r-1)}{k-1} & |s|=0 \text{ かつ } |s|<r \text{ の時,} \\ 1+\sum_{r'=1}^{|s|}\left\{k^{r'}\sum_{r_1=0}^{r'-1}C(|s|,r_1)\left(\frac{k-1}{k}\right)^{r_1} \right. & \\ \qquad \left. +\sum_{r_1=0}^{r'}C(|s|-r'+r_1,r_1)(k-1)^{r_1}\right\} & \\ \qquad +\left(2-\frac{1}{k}\right)^{|s|}\frac{k(k^r-k^{|s|})}{k-1} & |s|\ge 1 \text{ かつ } |s|<r \text{ の時.} \end{cases} \tag{2}$$

$$L_1 = 1+\sum_{r'=1}^{r}\left\{k^{r'}\sum_{r_1=0}^{r'-1}C(|s|,r_1)\left(\frac{k-1}{k}\right)^{r_1}+\sum_{r_1=0}^{r'}C(|s|-r'+r_1,r_1)(k-1)^{r_1}\right\}, \tag{3}$$

$$U_1 = \sum_{r'=0}^{r}\sum_{r_1=0}^{r'}\sum_{r_2=0}^{r'-r_1}C(|s|,r_1)(k-1)^{r_1}C(|s|-r_1,r_2)I(|s|-r_2,r'-r_1-r_2), \tag{4}$$

$$L_2 = 1+\left(2-\frac{1}{k}\right)^{|s|}\frac{k(k^r-1)}{k-1}, \tag{5}$$

$$U_2 = \sum_{r'=0}^{r}\sum_{r_1=0}^{|s|}\sum_{r_2=0}^{|s|-r_1}C(|s|,r_1)(k-1)^{r_1}C(|s|-r_1,r_2)I(|s|-r_2,r'-r_1-r_2), \tag{6}$$

$$L_3 = 1+\sum_{r'=1}^{|s|}\left\{k^{r'}\sum_{r_1=0}^{r'-1}C(|s|,r_1)\left(\frac{k-1}{k}\right)^{r_1}+\sum_{r_1=0}^{r'}C(|s|-r'+r_1,r_1)(k-1)^{r_1}\right\} + \left(2-\frac{1}{k}\right)^{|s|}\frac{k(k^r-k^{|s|})}{k-1}, \tag{7}$$

ここで,

$$I(|s|-r_2,r-r_1-r_2) = \begin{cases} 1 & r-r_1-r_2=0 \text{ の時,} \\ \sum_{q=1}^{r-r_1-r_2}C(|s|-r_2+1,q)C(r-r_1-r_2-1,q-1)k^{r-r_1-r_2} & \\ \qquad r-r_1-r_2\ge 1 \text{ かつ } r-r_1\le |s|+1 \text{ の時,} & \\ \sum_{q=1}^{|s|-r_2+1}C(|s|-r_2+1,q)C(r-r_1-r_2-1,q-1)k^{r-r_1-r_2} & \\ \qquad r-r_1-r_2\ge 1 \text{ かつ } r-r_1>|s|+1 \text{ の時.} & \end{cases}$$

に関して $\Theta(poly(|s|))$ に属し, (iii) k と $|s|$ が固定されている時, r に関して $\Theta(exp(r))$ に属する.

次に, A^* 上に Levenshtein 距離 d_L が定義されている場合の文字列球の体積の成長速度に関する結果を述べる. 任意の $k \in \mathbb{Z}^+, s \in A^*$, 及び $r \in \mathbb{N}$ に対して, (1) $1 \leq r \leq |s|$ の時, $L_1 \leq |U_{k,d_L}(s,r)| \leq U_1$, (2) $|s| = 0$ かつ $|s| < r$ の時, $L_2 \leq |U_{k,d_L}(s,r)| \leq U_2$, 及び (3) $|s| \geq 1$ かつ $|s| < r$ の時, $L_3 \leq |U_{k,d_L}(s,r)| \leq U_2$ が成り立つ. L_1, U_1, L_2, U_2, 及び L_3 はそれぞれ式 (3), (4), (5), (6), 及び (7) によって与えられる. 従って, $|U_{k,d_L}(s,r)|$ は, (i) $|s|$ と r が固定されている時, k に関して $\Theta(poly(k))$ に属し, (ii) k と r が固定されている時, $|s|$ に関して $\Theta(poly(|s|))$ に属し, (iii) k と $|s|$ が固定されている時, r に関して $\Omega(exp(r))$ と $O(poly(r) \times exp(r))$ に属する.

4. まとめ

$\mathbb{R}^n$ における超球と超球面の体積の公式は古くから知られており, L^2 ノルムのみならず, 一般の L^p ノルムの下での体積の公式も知られている [2]. しかし, これまで, どの $d \in D$ に対しても, A^* における球や球面の大きさの公式は知られていなかった.

指数分布の確率密度関数 $f(x;\lambda) = \lambda \exp(-\lambda x), x \geq 0, \lambda > 0$ の基準化定数は, 定積分

$$\int_0^\infty \exp(-\lambda x) \mathrm{d}x = \frac{1}{\lambda}$$

から容易に求められる. 一方で, 正規分布の確率密度関数 $f(x;\mu,\sigma^2) = 1/(\sqrt{2\pi}\sigma)\exp(-(x-\mu)^2/(2\sigma^2)), x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma^2 > 0$ の基準化定数は, 指数分布のそれほど簡単には求められず, Gauss 積分の公式

$$\int_{-\infty}^\infty \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x = \sqrt{2\pi}\sigma$$

を必要とする. A^* 上の Laplace 様分布の確率関数の基準化定数に関しても状況は似ており,

$$\sum_{s \in A^*} \left(\frac{\rho}{\rho+1}\right)^{d(s,\lambda)} = (\rho+1)|V_{k,d}(\lambda, d(s,\lambda))|$$

より, それが $1/((\rho+1)|V_{k,d}(\lambda, d(s,\lambda))|)$ という形を持つことはすぐに知られるが, これを具体的に計算するには, 文字列球面の大きさ $|V_{k,d}(\lambda, d(s,\lambda))|$ を計算する必要がある. 文字列球の大きさ $|U_{k,d}(\lambda, d(s,\lambda))|$ を計算できれば, $|V_{k,d}(\lambda, d(s,\lambda))|$ も計算でき, この逆ももちろん成り立つ.

文字列球は 1 つの正則言語であるため, その大きさを計算することは, 正則言語の大きさを計算することでもある. 理論と応用の両方の重要性から, 特に d_L の下での文字列球の大きさの明示公式を求める問題は極めて興味深いが, 未解決である (Becerra–Bonache *et al.* [1] 参照). 本研究では, $d_{H'}$ の下での文字列球の大きさの明示的な公式の導出には成功したが, d_L の下での文字列球の大きさの公式を得るには至らず, その成長速度を評価するに止まった. しかし, d_L の下での文字列球の真の大きさと 0.99 以上の確率で等しい推定値を返す効率的な乱択アルゴリズムを構成し, それを用いて数値実験を行うことにより, その公式に関する 1 つの予想を得た. 発表では, その乱択アルゴリズムとそれから得られた予想についても述べる. 本発表は [3] に基づいている.

参考文献

[1] Becerra-Bonache, L., De La Higuera, C., Janodet, J.-C. and Tantini, F.: Learning balls of strings from edit corrections, *J. Mach. Learn. Res.*, Vol. 9, pp. 1841–1870 (2008).
[2] Dirichlet, P. G. L.: Sur une nouvelle méthode pour la détermination des intégrales multiples, *J. Math. Pures Appl.*, Vol. 4, pp. 164–168 (1839).
[3] Koyano, H. and Hayashida, M.: Growth rates and sizes of balls of strings under the extended Hamming and Levenshtein distances, submitted.
[4] Koyano, H., Hayashida, M. and Akutsu, T.: Maximum margin classifier working in a set of strings, *Proceedings of the Royal Society A*, Vol. 472, No. 2187, 20150551 (2016).
[5] 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*, Vol. 106, pp. 94–128 (2019).
[6] Koyano, H. and Kishino, H.: Quantifying biodiversity and asymptotics for a sequence of random strings, *Physical Review E*, Vol. 81, No. 6, 061912 (2010).
[7] Koyano, H., Sawada, K., Yamamoto, N. and Yamada, T.: Modeling and analysis of the dynamics of microbial communities in environments, submitted.
[8] Koyano, H., Tsubouchi, T., Kishino, H. and Akutsu, T.: Archaeal β diversity patterns under the seafloor along geochemical gradients, *Journal of Geophysical Research G: Biogeosciences*, Vol. 119, No. 9, pp. 1770–1788 (2014).
[9] 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].