修士論文 要旨

環境を介して情報共有する細胞集団の強化学習的理解

数理情報学専攻 48226209　加藤 雅己
指導教員　小林 徹也 教授

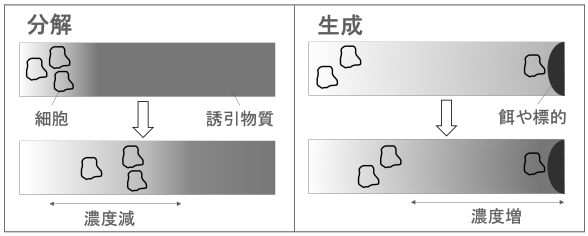

図 1. 細胞集団による誘引物質の分解・生成の模式図.

1　背景

環境を介した情報共有を伴う集団移動　集団移動は形態形成から採餌に至るまで生物系において普遍的に見られる問題である．特徴的な系として，餌などから発せられる誘引物質 (≈ 匂い) の勾配をただ登るだけでなく，自ら誘引物質を分解・生成し濃度勾配を能動的に変化させることで情報を間接的に共有し，効率的な移動と集積を実現する細胞集団が挙げられる [1, 2] (図 1)．個々の細胞が周囲の環境に応じて化学物質を分解・生成し，その量を調整しながらことで情報共有する仕組みは免疫細胞や癌細胞など細胞集団に共通して見られ，単純な感知機構しか有していない小さな個体の集団が，広大な環境中の情報を効率よく共有し高度な移動能を実現する上で重要な役割を果たしていると考えられる．

情報共有する細胞集団の規範的モデル化　誘引物質の分解・生成は，古くは現象論的に生成項と分解項を導入した反応拡散方程式で記述され，近年は分子生物学的な知見を活用したモデルの構築も進められている [3]．しかしながら，生成や分解がどういう目的を持っているのか，定常状態の濃度場にどういう意味があるのかといった機能的側面は明らかではない．

機能の観点から生体情報処理のモデルを構築する手法が規範的モデル化であり，脳の学習や大腸菌の化学走性や免疫系などの様々な生体情報処理で応用されている．規範的なモデル化では，生存に重要な戦略は進化の過程の選択圧によって生存に必要な目的に対して最適化されているという考えの元，最適化理論を用いてモデルが構築される．構築されたモデルは現象や他手法によるモデルと比較され，現象論や分子生物学的知見だけからは分からなかった関係性や解釈を見出している．

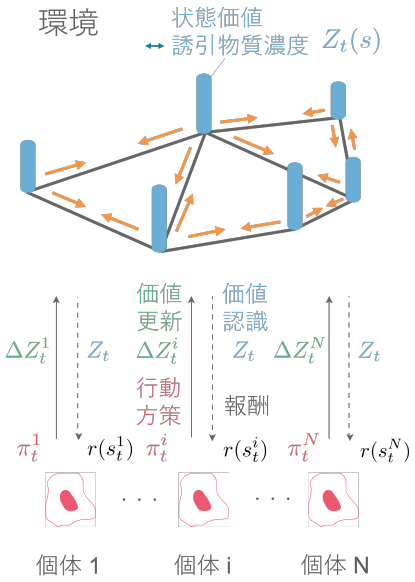

図 2. 経験に基づき協調的に状態価値を学習する個体集団モデルの模式図．環境 $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ の頂点上に状態価値 $Z_t : \mathcal{S} \to \mathbb{R}_{\geq 0}$ が定義され，各個体 $i \in \{1, \ldots, N\}$ は状態 $s_t^i \in \mathcal{S}$ に応じて環境から個別に報酬 $r(s_t^i) \in \mathbb{R}$ を受け取り，自分の周囲の状態価値 Z_t を認識する．そして，報酬 $r(s_t^i)$ と状態価値 Z_t に基づいて行動方策 $\pi_t^i(\cdot|s_t^i) \in \Delta(\mathcal{S})$ (2) と更新差分 $\Delta Z_t^i \in \mathbb{R}$ (4c) を計算し，移動と状態価値の更新を行う．

しかし，細胞集団の情報共有を伴う移動戦略については規範的なモデル化はほとんど確立されていない．

2　研究目的・方法

研究目的　本研究では，誘引化学物質を分解・生成しながら探索を行う細胞集団を情報処理システムとして捉え，規範的なモデル化を行う．そして現象論的モデルや実験事実と比較することで，なぜ細胞集団は化学物質を介して情報共有をするのかについて示唆を与える．

研究方法　細胞は餌などの目標に到達しやすくなるように目的を持って誘引物質を分解・生成しているであろう，という仮定に基づいて誘引物質の分解・生成アルゴリズムを分散的強化学習の観点から構築した．

環境中を探索する個体集団のダイナミクスは，環境を表す無向グラフ $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ 及び，$i \in \{1, \ldots, N(\in \mathbb{Z}_+)\}$; $t = 0, 1, \ldots$; $s \in \mathcal{S}$ に対して N 個体の状態 $s_t^i \in \mathcal{S}$，共通の行動方策 $\pi_t^i(\cdot|s) = \pi_t^*(\cdot|s) \in \Delta(\mathcal{S})$，共

通の報酬関数 $r(s) \in \mathbb{R}$ の 4 つの要素からなるマルコフ決定過程で記述される．

通常の強化学習では**個体内部に定義される状態価値を環境中を拡散する誘引物質 Z_t と同一視する**ことで [4]，拡散で揺らぐ誘引物質を自ら分解・生成しながら濃度勾配を登る細胞集団を，状態価値 Z_t が最適価値 $Z^*(s) =:= \exp(V^*(s))$ $(V^*(s) := \max_\pi V^\pi(s))$ [5] に近づくよう一定の正則化のもとで分散的に更新しつつ状態価値の勾配を登る個体集団としてモデル化した．行動方策 π_t^* は，状態価値 $Z_t = Z^*$ のとき最適方策 $\pi_t^* = \pi^* := \operatorname{argmax} V^\pi$ となる貪欲方策として定める：

$$s_{t+1}^i \sim \pi_t^*(\cdot|s_t^i) := \frac{p(\cdot|s_t^i)Z_t^\gamma(\cdot)}{\sum_{s'\in\mathcal{S}} p(s'|s_t^i)Z_t^\gamma(s')}, \tag{1}$$

$$V^\pi(s) := E\left[\sum_{t=0}^{\infty}\gamma^t R^\pi(s_t, s_{t+1})\middle| s_0 = s\right], \tag{2}$$

$$R^\pi(s_t, s_{t+1}) = r(s_t) - \log\frac{\pi(s_{t+1}|s_t)}{p(s_{t+1}|s_t)}. \tag{3}$$

ただし，$p : \mathcal{S}\times\mathcal{S} \to [0,1]$ はランダムウォークのような個体に内在する定常方策である．また，細胞集団は (進化の結果として) 正則化付き推定二乗誤差 $\mathcal{C}[\boldsymbol{Z}_t]$ (略) を最小化していると仮定し，状態価値の更新アルゴリズムは $\mathcal{C}[\boldsymbol{Z}_t]$ の最急降下の近似として定める：

$$Z_{t+1}(s) - Z_t(s) = -\sum_{i=1}^{N}\mathbb{1}_{s_t^i=s}\alpha\Delta Z_t(s) \tag{4a}$$

$$- D\sum_{s'\in\mathcal{S}} L_{ss'}Z_t(s'), \tag{4b}$$

$$\Delta Z_t(s) := Z_t(s) - \exp(r(s))\sum_{s'\in\mathcal{S}} p(s'|s)Z_t^\gamma(s'). \tag{4c}$$

ここで，$D \in \mathbb{R}_+$ を正則化係数，$\alpha \in (0,1)$ をステップサイズ，L を環境 $\mathcal{G}$ のグラフラプラシアン，$\mathbb{1}_{s_t^i=s}$ は個体 i が時刻 t に頂点 s にいる場合 1，いない場合 0 を取る指示関数とする．式 (4) では，各個体が自身が獲得した報酬と周囲の濃度から計算される近似的な推定誤差から滞在位置 s の濃度の更新差分である $\Delta Z_t(s)$ を環境に出力し，各時刻で濃度は各個体の更新差分の和だけ協調的に更新される (4a)．その一方で，環境全体では誘引物質はある点に留まることなく個体とは独立な一定の強度で周囲に拡散する (4b) というモデルになっている．

提案したモデルは，以下のような連続時間・連続空間・無限個体極限を持つ：

$$\partial_t\mu(t,x) = -\nabla\cdot(\gamma\sigma^2\mu(t,x)\nabla\ln Z(t,x)) + \frac{\sigma^2}{2}\nabla^2\mu(t,x), \tag{5a}$$

$$\partial_t Z(t,x) = -\alpha\mu(t,x)Z(t,x) + \alpha\mu(t,x)\exp(r(x))Z^\gamma(t,x) + D\sigma^2\nabla^2 Z(t,x). \tag{5b}$$

ここで，μ は個体密度，Z は誘引物質の濃度，$\sigma \in \mathbb{R}_{\geq 0}$ は個体の拡散係数である．式 (5) は現象論的モデル [3, 4] と式の形がほとんど対応し，細胞集団は誘引物質の分解・生成を通じて目標位置を分散的に学習していることが示唆される．また，現象論的モデルでは生成項と分解項をヒューリスティックに定めていたが，本研究では最適性の観点から学習ダイナミクスとなるような生成項と分解項の一つを導いている．

最後に，生物的な問題設定で数値実験を行いモデルを評価した．まず，複雑な環境であっても学習可能であること，学習ダイナミクスが集団としても個体としても実験事実と定性的に整合することが確認された．そして，通常の強化学習アルゴリズムにはない状態価値の正則化 (拡散) が，報酬の累積和という元の目的を悪化させるもののある種の移動性能を高める正則化として，またより大きな報酬への局在を促進する特殊な平滑化として機能していることが発見された．この結果は，細胞集団は環境中を揺らぐ誘引物質を介して目標に関する情報を共有することで，化学走性戦略を変えることなく探索性能や移動性能を高めていることを示唆する．

参考文献

[1] Robert H. Insall, Peggy Paschke, and Luke Tweedy. Steering yourself by the bootstraps: how cells create their own gradients for chemotaxis. Trends in Cell Biology, Vol. 32, No. 7, pp. 585–596, Jul 2022.

[2] Katharina M Glaser, Michael Mihlan, and Tim Lämmermann. Positive feedback amplification in swarming immune cell populations. Current Opinion in Cell Biology, Vol. 72, pp. 156–162, 2021.

[3] Kevin J. Painter. Mathematical models for chemotaxis and their applications in self-organisation phenomena. Journal of Theoretical Biology, Vol. 481, pp. 162–182, 2019.

[4] Alberto Pezzotta, Matteo Adorisio, and Antonio Celani. Chemotaxis emerges as the optimal solution to cooperative search games. Physical Review E, Vol. 98, No. 4, Oct 2018.

[5] Emanuel Todorov. Efficient computation of optimal actions. Proceedings of the National Academy of Sciences - PNAS, Vol. 106, No. 28, pp. 11478–11483, Jul 2009.