修士論文 要旨

# 適応免疫系の強化学習的理解

数理情報学専攻 48-166208 **加藤卓也**

指導教員 **小林徹也 准教授**

## 1 はじめに

適応免疫系の特徴は、病原体それぞれに適した攻撃細胞の応答パターンを学習できる点にある。このような学習は適応免疫系の細胞集団の中でも主に T ヘルパー細胞集団と呼ばれる細胞集団によって担われている。本論文ではこの T ヘルパー細胞集団の学習ダイナミクスの数理モデルの構築を試みる。

## 2 学習ダイナミクスの強化学習との対応

T ヘルパー細胞集団の学習は、各病原体に対して様々な応答パターンを試すことでそれらがその病原体にどれほど効果的であるかを知り、効果が比較的大きい応答パターンを記憶していく、試行錯誤のプロセスとして記述することができる。このような描像に従えば、免疫細胞集団のダイナミクスは以下の特徴で記述できる。下記の 4 つの特徴は強化学習におけるマルコフ決定過程 (MDP) の 4 つの特徴、状態集合、行動集合、遷移確率、報酬とそれぞれ対応する。

- 抗原パターン集合 $\mathcal{S} = \{0,1\}^N$。病原体の一部である抗原の内、どの種類のものが存在するかを示す。$N$ 種類の存在が仮定され、種類 $i$ の抗原が存在する場合 $s_i = 1$ と表現。
- 攻撃細胞の活性化パターン集合 $\mathcal{A} = \{0,1\}^M$。$M$ 種類の攻撃細胞の存在が仮定され、種類 $j$ の攻撃細胞が活性化していれば $a_j = 1$ と表現。
- 抗原パターンが $s$ であるときに攻撃細胞の活性化パターンが $a$ となった際、その影響により抗原分布が $s'$ となる確率 $P(s'|s,a)$。
- 抗原分布が $s$ であるときに攻撃細胞の活性化分布 $a$ がどれほど効果的かを表す報酬 $r = R(s,a) \in \mathbb{R}$。効果が高いときに大きい値を取ると仮定。

T ヘルパー細胞集団は抗原パターン $s$ の情報を元に攻撃細胞へと攻撃の指示をだし、彼らの活性化パターン $a$ を定める役割を担う。これを抗原パターン $s$ で条件づけられた攻撃細胞の活性化パターン $a$ 上の確率分布 $\pi(a|s)$ を定めることで表現し、戦略と呼ぶ。彼らはより高い報酬 $R(s,a)$ を得られるように戦略 $\pi(a|s)$ を変化させる。

## 3 T ヘルパー細胞集団のモデル

T ヘルパー細胞集団の数理モデルをいくつかの既存の強化学習モデルを発展させることで構築する。

### 3.1 RBM を発展させた T ヘルパー細胞集団モデル

Restricted Boltzmann Machine (RBM) を使用した強化学習モデル [1] を応用して数理モデルを構築していく。今、T ヘルパー細胞は $K$ 種類存在し、各種類 $k$ の細胞は $n_k$ 個存在すると仮定する。また、種類 $k$ の $l$ 個目の細胞の活性度を $h_l^k \in \{0,1\}$ と表現し、彼らは抗原 $i$ と $w_{ki} \in \mathbb{R}$ の親和性をもち、攻撃細胞 $j$ に $u_{kj} \in \mathbb{R}$ の影響を持つと仮定する。このことを RBM の記述に準じ、以下のようなハミルトニアンで表す。

$$H(s,a,h) = \sum_{k=1}^{K}\sum_{l=1}^{n_k} h_l^k \Big(\sum_{i=1}^{N} w_{ki}s_i + \sum_{j=1}^{M} u_{kj}a_j\Big), \quad (1)$$

細胞集団の戦略 $\pi(a|s)$ はこのハミルトニアンにより定義されるボルツマン分布に従うと仮定し

$$\pi(a|s) = \frac{\sum_h \exp(-H(s,a,h))}{\sum_{h,a} \exp(-H(s,a,h))}, \quad (2)$$

と表す。以上の戦略に SARSA 学習を適用すると、細胞数分布の変化はエピソード $\{s,a,r,s',a'\}$ に対して

$$n_k \leftarrow n_k + \alpha n_k \lambda_k(s,a,s',a'), \quad (3)$$

となる。ただしここで $\alpha$ は学習率である。さらに $\lambda_k$ は各種類 $k$ の T ヘルパー細胞の増殖率であり

$$\lambda_k = G_k(s,a) + E_k(s,a,s',a') - D_k(s,a), \quad (4)$$
$$G_k = r\log(1+\exp(\tau_k(s,a))), \quad (5)$$
$$E_k = \gamma\tilde{Q}(s',a')\log(1+\exp(\tau_k(s,a))), \quad (6)$$
$$D_k = \tilde{Q}(s,a)\log(1+\exp(\tau_k(s,a))), \quad (7)$$
$$\tau_k = \sum\nolimits_i w_{ki}s_i + \sum\nolimits_j u_{kj}a_j, \quad (8)$$

となっている。

導出された学習ダイナミクスは現実の T ヘルパー細胞と多くの類似点を持つ。$G_k$ の 2 つの項はそれぞれ樹状細胞表面の補助刺激分子による T 細胞表面のレセプター CD28 への刺激 ($r$ に対応) と抗原による TCR への刺激 ($\log(1+\exp(\tau_k))$ に対応) と考えることができる。この解釈の下では、どのようにして T ヘルパー細胞集団の学習に寄与しているかが不明であった CD28

への補助刺激分子による刺激 [2] は、T ヘルパー細胞集団に対する報酬の役割を持っている可能性が示唆される。また $D_k$ に対応する現象として Fas リガンドによる活性化誘導アポトーシスがある。この対応が正しければ、単に細胞数が増えすぎないように T ヘルパー細胞を減らすためだけの機能であるとされていた活性化誘導アポトーシス [3] は、このモデルにおいて細胞を選択的に減らすことでその分布を調整し、学習を進ませる役割が示唆される。

しかしこのモデルにおいては、学習ダイナミクスのうち $E_k$ の効果が生物的な解釈が難しく、さらに戦略 $\pi(a|s)$ の生物学的対応が不明確である。

### 3.2 DQN を発展させた T ヘルパー細胞集団モデル

以下では Deep Q Network (DQN)[4] を応用し、戦略 $\pi(a|s)$ の生物的な解釈が容易であるモデルを提案する。今、種類 $k$ の T ヘルパー細胞部分集団は抗原分布 $s$ によって

$$h_k(s) = \sigma(\sum\nolimits_i w_{ki} s_i), \tag{9}$$

だけ活性化されると仮定する。ただしここで $\sigma$ はシグモイド関数である。また、種類 $j$ の攻撃細胞の活性度 $a_j$ は T ヘルパー細胞の活性度に依存して確率的に

$$p(a_j = 1|s) = \sigma(\sum\nolimits_k u_{jk} n_k h_k(s)), \tag{10}$$

で定まるとする。よって T ヘルパー細胞集団の戦略は

$$\pi(a|s) = \frac{\exp(\tilde{Q}(s,a))}{\sum_a \exp(\tilde{Q}(s,a))}, \tag{11}$$

$$\tilde{Q}(s,a) = \sum\nolimits_k n_k h_k(s) \sum\nolimits_j u_{jk} a_j, \tag{12}$$

となる。RBM を応用したモデルにおいては戦略 $\pi$ の生物解釈が不明確であったが、このモデルはそのような問題を持たない。ここに SARSA 学習アルゴリズムを当てはめると、エピソード $\{s, a, r, s', a'\}$ に際して細胞数分布の変化は

$$n_k \leftarrow n_k + \alpha n_k \lambda_k(s, a, s', a'), \tag{13}$$

となり、増殖率 $\lambda_k$ は

$$\lambda_k = G_k(s,a) + E_k(s,a,s',a') - D_k(s,a), \tag{14}$$

$$G_k = r h_k(s) \Sigma_j u_{kj} a_j, \tag{15}$$

$$E_k = \gamma \tilde{Q}(s',a') h_k(s) \Sigma_j u_{kj} a_j, \tag{16}$$

$$D_k = \tilde{Q}(s,a) h_k(s) \Sigma_j u_{kj} a_j, \tag{17}$$

となる。これらの項の内 $G_k$ と $D_k$ は RBM モデルと同様の生物学的な解釈が可能である。しかしこのモデルにおいても、学習ダイナミクスのうち $E_k$ の効果の生物的な解釈が難しい。

この項 $E_k$ の生物学的な解釈を容易にするため、確率的山登り法 [5] を発展させた学習ダイナミクスを提案する。このダイナミクスにおいて細胞集団分布 $\{n_k\}_{k=1}^K$ の変化は

$$z_k \leftarrow \gamma z_k + \alpha h_k(s) \sum\nolimits_j u_{kj}(a_j - \bar{a}_j) \tag{18}$$

$$n_k \leftarrow n_k + \alpha n_k r z_k, \tag{19}$$

$$\bar{a}_j = \sigma(\sum\nolimits_k u_{jk} n_k h_k(s)), \tag{20}$$

と記述され、項 $E_k$ はでてこない。ただしここで $\alpha$ は学習率である。各種類 $k$ の T ヘルパー細胞が外部からの刺激を $z_k$ に記憶して、その値に依存して増殖するような描像となっている。$z_k$ が細胞内分子の濃度などに対応すると考えればこのようなダイナミクスは生物的な解釈が容易であり、T ヘルパー細胞について似たダイナミクスは確かめられている [6]。

## 4 結論

以上で提案した学習ダイナミクスはそれぞれ現実の T ヘルパー細胞と類似する点を多く持っている。それらの類似点は T ヘルパー細胞において知られた性質である、樹状細胞からの補助刺激、活性化誘導アポトーシスなどが学習に対してどのように寄与しているかを示唆することができる。

## 参考文献


[1] B. Sallans et al., Reinforcement learning with factored states and actions, J. Mach. Learn. Res., Vol. 5, August, pp. 1063-1088, 2004

[2] L. Chen et al., Molecular mechanisms of T cell co-stimulation and co-inhibition, Nat. Rev. Immunol. 13(4), 337-242, 2013

[3] S. Maher et al., Activation-induced cell death: the controversial role of Fas and Fas ligand in immune privilege and tumour counterattack, Immunology and cell biology, 80(2), 131-137, 2002

[4] V. Mnih et al., Playing atari with deep reinforcement learning, arXiv preprint arXiv:1312.5602, 2013

[5] H. Kimura et al., Reinforcement learning by stochastic hill climbing on discounted reward, ICML 1995

[6] J. Rachmilewitz et al., A temporal and spatial summation model for T-cell activation: signal integration and antigen decoding, 23(12), 592-595, 2002