論　文

神経活動の解析に基づく腹側線条体の強化学習機能のモデル化

篠塚　正成[†*a)]　　森田　昌彦[††b)]　　設楽　宗孝[†††]

Modeling the Function of the Ventral Striatum in Reinforcement Learning Based on the Analysis of Neuronal Activity

Masanari SHINOTSUKA[†*a)], Masahiko MORITA[††b)], and Munetaka SHIDARA[†††]

あらまし　大脳基底核で TD 学習が行われているという生理学的知見に基づいて，脳の強化学習モデルが幾つか提案されている．そのほとんどは線条体の striosome という領域が状態価値を表現するものとしているが，それ以外の可能性については十分に検討されていない．本研究では，striosome を多く含む腹側線条体に関する生理データを見直すことによって，強化学習における線条体の機能に関して新たなモデルを提案する．まず，視覚キュー付き報酬課題を学習したサルの腹側線条体において観測された報酬予測的なニューロン活動のデータを再解析したところ，これらは予測される報酬よりもむしろ過去の刺激や報酬の履歴を反映していることがわかった．そこで，「腹側線条体は，過去の履歴から状態価値を推定するのに適した中間表現を保持している」という仮説を立て，刺激や報酬の時系列からそのような表現を獲得可能な神経回路モデルを構築した．計算機シミュレーションの結果，このモデルにより再解析で見られたさまざまなニューロン活動のパターンが再現されることがわかった．このことは，腹側線条体にそのような機能があることを示唆するとともに，大脳基底核における強化学習が効率的な学習に適した状態空間の構成と並行して行われている可能性を示しており，強化学習を工学的に応用する上でも有意義である．

キーワード　大脳基底核，腹側線条体，強化学習，状態価値関数，リカレントニューラルネット

1. まえがき

機械学習の枠組みの一つとして強化学習が知られている [1]．強化学習は心理学において研究されていた動物の試行錯誤的学習を工学の最適制御の理論と結び付けて定式化したものであるが，近年強化学習と脳の神経回路の関連性が指摘されている．Schultz らは大脳基底核という脳領域内に存在するドーパミンニューロンが強化学習における TD 誤差に相当する活動を示すことを発見した [2]．これを根拠として，大脳基底核は強化学習を行う神経回路であるという仮説が提案され，様々なモデル化が試みられている．

その代表的なものとして，Barto のモデル [3] や Doya のモデル [4] があるが，これらのモデルでは，いずれも線条体の striosome が状態価値 $V(s)$，すなわち将来報酬の期待値を表現するとしている．この仮定は現在広く受け入れられているが，「状態価値を直接表現することが striosome の役割である」と言い切ることはできない．それ以外の役割については十分に検討されていないし，後述のように状態価値がどのような形で表現されているかについても議論があるからである．

最近，striosome の領域を多く含む腹側線条体において，過去の履歴を反映したニューロン活動が報告されている [5], [6]．将来の報酬が過去の刺激や報酬から予測可能である場合，履歴と将来報酬とを区別するのは難しい．そのため，これまで将来の報酬を反映しているとされてきたニューロン活動の中にも，過去の履歴が反映されている可能性がある．もしそうであるな

† 筑波大学大学院システム情報工学研究科，つくば市
Graduate School of Systems and Information Engineering, University of Tsukuba, 1–1–1 Tennodai, Tsukuba-shi, 305–8573 Japan
†† 筑波大学システム情報系，つくば市
Faculty of Engineering, Information and Systems, University of Tsukuba, 1–1–1 Tennodai, Tsukuba-shi, 305–8573 Japan
††† 筑波大学医学医療系，つくば市
Faculty of Medicine, University of Tsukuba, 1–1–1 Tennodai, Tsukuba-shi, 305–8577 Japan
* 現在，日本ユニシス株式会社
a) E-mail: m.shinotsuka2@gmail.com
b) E-mail: mor@bcl.esys.tsukuba.ac.jp
DOI:10.14923/transinfj.2014JDP7137

らば，それが強化学習において果たす役割を計算論的に考え，モデルに取り入れるべきであろう．

このような観点から，本研究では，腹側線条体のニューロン活動に関する Shidara ら [7] の実験データを再解析し，過去の履歴が反映されていることを示す．また，解析結果に基づいて腹側線条体の機能に関する新たな機能に関する仮説を提案し，その機能をモデル化する．更に，計算機シミュレーションの結果を生理データと比較することによって，モデルの妥当性を検証する．

2. 背　　景

2.1 強化学習と大脳基底核

2.1.1 状態価値関数

強化学習の主要な目的の一つは「状態価値」の推定である．必要な状態価値を全て正しく推定できれば，それを基に最適な行動を獲得することができる．本研究では，状態価値の推定に焦点を絞り，学習が行動に依存しない（報酬が行動に関係なく決まる）状況を考えることにする．

強化学習において，状態価値は「ある状態の後に得られる報酬の総和の期待値」と定義される．数式では，時刻 t における状態 s_t における状態価値を

$$V(s_t) = E[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots] \qquad (1)$$

と表し，状態価値関数と呼ぶ．この式で，r_t は時刻 t で与えられる報酬，E は期待値を表す．また，γ は 0 から 1 の実数値をとるパラメータで，割引率と呼ばれる．遠い将来の報酬ほど大きく割り引かれ，近い将来の報酬ほど状態価値に大きな影響を与えるため，一般に $V(s_t)$ は報酬が得られる状態に近づくにつれて増加する．

2.1.2 TD 誤 差

よく用いられる強化学習のアルゴリズムでは，状態遷移をするたびに

$$V_{\mathrm{new}}(s_{t-1}) \leftarrow V_{\mathrm{old}}(s_{t-1}) + \alpha\delta_{t-1} \qquad (2)$$

という式を用いて状態価値関数を更新する．ここで δ_{t-1} は

$$\delta_{t-1} = r_t + \gamma V(s_t) - V(s_{t-1}) \qquad (3)$$

によって与えられる量であり，TD（temporal difference）誤差と呼ばれる．TD 誤差は，学習が不十分なうちは大きな値をとって状態価値関数を大きく修正するが，学習が進行し状態価値関数が正しく推定できるようになってくると，式 (1)，(3) より 0 に収束していく．このことから，TD 誤差には「予想外の報酬（無報酬）に対して正（負）の値を示す」という性質がある．

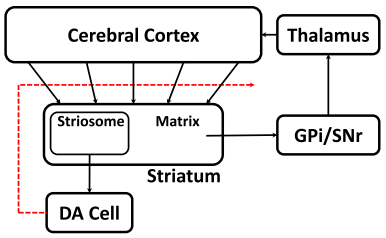


図 1　大脳基底核の神経回路
Fig. 1　Neural circuits of the basal ganglia.

2.1.3 大脳基底核の構造

大脳基底核（basal ganglia）は大脳皮質（cerebral cortex）に包まれるように存在する特定の脳部位の総称である．大脳基底核は大脳皮質からの入力を受け，視床を介して大脳皮質へ出力をフィードバックする，というループ回路を形成する．このループ回路は入力となる大脳皮質の領野ごとに閉回路を形成し，辺縁系皮質を結ぶ「辺縁系ループ」や運動野を結ぶ「運動ループ」などが存在する．

図 1 は大脳基底核の主な神経回路を簡略化した模式図である．入力部である線条体 (striatum) は大脳皮質の広範囲から視覚や聴覚等の情報を受ける．また線条体は，striosome と matrix という二つの領域で構成され，striosome は中脳ドーパミンニューロン (DA cell) に出力を送り，matrix は淡蒼球内節 (internal segment of globus pallidus, GPi) と黒質網様部 (substantia nigra pars reticulata, SNr) に出力を送る．中脳ドーパミンニューロンは出力を線条体にフィードバックし，GPi/SNr は視床 (thalamus) を介して出力を大脳皮質にフィードバックする．

2.1.4 中脳ドーパミンニューロン

Schultz ら [2] はサルに視覚刺激と報酬の条件付け課題を行わせ，その際のドーパミンニューロンの活動を記録した．最初ドーパミンニューロンは報酬に対して興奮性の応答を示したが，学習後は報酬に対する直接の応答は消え，条件刺激の提示時に応答が現れるようになった．更に，本来報酬が与えられるべきタイミングで報酬が与えられないと，その時刻に抑制性の応答が現れた．つまり，ドーパミンニューロンは予想外に都合の良い刺激（学習前の報酬そのものや学習後の報

酬予告刺激）に対し応答を増加させ，予想外に都合の悪い刺激（学習後における報酬が得られるべきタイミングの無報酬）に対して応答を減少させる．これらの特徴が強化学習における TD 誤差の特徴と一致することから，大脳基底核は強化学習を行う神経回路であるという仮説の有力な根拠となっている．

2.1.5 線　条　体

線条体は大脳基底核の入力部に位置している．線条体には striosome，matrix という神経連絡関係の異なる二つのコンパートメントが存在し，striosome はドーパミンニューロンに，matrix は淡蒼球に出力を送ることが知られている．また，これとは別に腹側線条体，背側線条体といった解剖学的な分類がなされる場合もあり，腹側線条体には striosome の割合が多い [8]．

線条体ニューロンは，細胞外のドーパミン濃度に依存した学習を行うという仮説が有力である [9]．一般的なニューロンの学習則として，入力側と出力側が同時に発火したとき結合強度が強まるという Hebb 則が知られているが，線条体においては同時に細胞外のドーパミン濃度が通常より高くなっている必要があるとされている．つまり入力側の発火，出力側の発火，ドーパミン濃度の上昇の三つがそろったときに学習が行われる（濃度が低いときには逆に強度が弱まる）．この仮説が正しければ，線条体はドーパミン投射，すなわち TD 誤差信号の入力が豊富であるため，それを用いて状態価値や行動の学習を行うことが可能であると考えられる．

2.1.6 大脳基底核における強化学習のモデル

以上のような解剖学的，生理学的知見から，大脳基底核と強化学習の関連性は強いと考えられており，大脳基底核の強化学習モデルが複数提案されている．代表的なものとして，線条体の matrix と striosome がそれぞれ actor と critic に相当するという Barto のモデル [3] や，matrix が行動価値 $Q(s,a)$ を表現するという Doya のモデル [4] が挙げられる．

二つのモデルは，辺縁系大脳皮質からの入力が多い striosome において状態価値が学習されるとする点で共通している．具体的には，図 2 に示すように，striosome は辺縁系皮質から観測した現在状態 s_t を受け取り，状態価値 $V(s_t)$ を計算する．また中脳ドーパミンニューロンは striosome から状態価値を受け TD 誤差を計算し，striosome へフィードバックする．フィードバックされた TD 誤差を用いて striosome は状態価値関数の修正を行う，とする．

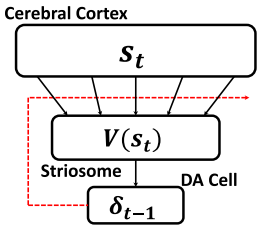


図 2　従来の大脳基底核の強化学習モデルに共通する構造
Fig. 2　Structure common to conventional reinforcement learning models of the basal ganglia.

striosome，または striosome を多く含む領域である腹側線条体が状態価値を表現するというモデルは，腹側線条体がドーパミンニューロンの投射を受けていること，報酬予測に関わることを示す知見 [10], [11] が多数あることなどから，広く支持されている．その一方で，このモデルに関して幾つかの疑問がある．

一つは，状態価値がどのような形で表現されているかという点である．もし，単純に単一ニューロンまたはニューロン集団の活動の大きさによって表現されているのならば，腹側線条体には状態価値を直接反映したニューロン活動が多く見られるはずである．

しかし，例えば Cromwell ら [11] の実験では，期待される報酬量の大中小に依存したニューロン活動が観測されたが，その中で報酬量が大きいほど活動が大きい（または小さい）という一貫した活動は少数であった．また，報酬量が同じであっても時間的な近さに応じて活動が変化するはずであるが，報酬に近づくにつれて活動が増える（または減る）といったニューロン活動が多数観測されたという報告はない．例えば，後述する Shidara ら [7] の実験では，報酬までの近さに依存した活動は見られたが，状態価値を直接表現するような単純なものではなかった．

こうしたことから，鮫島ら [12] は,「腹側線条体ニューロンは状態価値関数を直接表現しているのではなく，その基底関数を表現しており，投射先である淡蒼球などにおいて状態価値が合成される」という説を出している．

もう一つの疑問は，状態価値の表現以外の機能はもたないのか，あるいは状態価値以外の情報は表現されていないのか，というものである．これに関して，最近 Goldstein ら [5] は，ラットの腹側線条体ニューロン活動に，過去の報酬と予測報酬の両方が反映されていることを示した．同様に，Kim ら [6] は，前の試行

でラットが取った行動がニューロン活動に反映されていることを示した．これらの知見は，striosome の機能が状態価値を表現することだけではないことを示唆している．

2.2 腹側線条体のニューロン活動

ここでは，Shidara ら [7] の実験で用いた多試行報酬スケジュール課題の説明と実験結果，及びそれに対する従来の解釈について述べる．

2.2.1 多試行報酬スケジュール課題

まず，視覚弁別バーリリース課題（図 3 A）について説明する．初めにモニター上部に長方形のキュー（後述）が表示される．サルがバーを握ると，モニターに白の注視点が表示され，注視に成功すると赤（Wait）のターゲットが表示される．ターゲットが赤の間バーを把持し続け，ランダムな待ち時間の後ターゲットが緑（Go）に変わった際にバーを放すと，ターゲットが青（OK）に変わり報酬としてジュースが与えられる．以上の視覚弁別バーリリース課題 1 回を「試行」と呼ぶ．

この試行を 1〜3 回ずつ組にしたのが多試行報酬スケジュール課題（図 3 B）である．この試行の組のことを「スケジュール」と呼び，1 スケジュールの試行を全て成功した場合にのみ報酬が与えられる．例えばスケジュールが 3 試行からなる場合，1，2 試行目では成功しても報酬が与えられず，3 試行目まで連続して成功して初めて報酬が与えられる．

現在の試行が何試行スケジュールの何番目であるかを 1/2（2 試行のうちの 1 番目）や 2/3（3 試行中の 2 番目）といった分数で表し，スケジュール進行度と呼ぶ（図 3 B）．画面上部のキューは，この分数値に比例した明るさで表示される．一つのスケジュールが終わると次のスケジュールが三つの内からランダムに選ばれ，以後これを繰り返す．

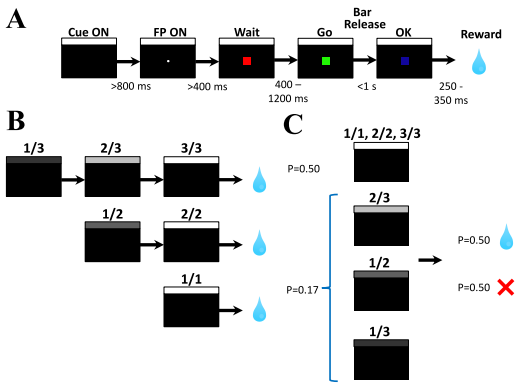


図 3 多試行報酬スケジュール課題（Shidara ら [7] より改変）

Fig. 3 Multiple trial reward schedule task (adapted from Shidara et al. [7]).

以上が実験条件であり，以下「キュー条件」と呼ぶ．このほかにコントロール条件として，提示キューと報酬の有無をランダムにした「ランダム条件」が設定されている（図 3 C）．ランダム条件では，全てのキュー（1/1, 1/2, 2/2, 1/3, 2/3, 3/3）が等確率で選ばれるが，1/1, 2/2, 3/3 は白色の同一キューであるため，実際には白色のキュー（キュー 1 と呼ぶ）が確率 1/2 で，1/2, 1/3, 2/3 のキューがそれぞれ確率 1/6 で提示される．また，報酬が与えられる確率は，提示キューにかかわらず一律 1/2 である．

2.2.2 測定結果と従来の解釈

課題（キュー条件）を十分に訓練したサルを用いて，腹側線条体ニューロンの活動を記録した．毎日のセッションではキュー条件を 100〜200 試行程度行い，その後計測対象のニューロンがまだ計測可能であれば，更に 100〜200 試行程度ランダム条件を行った．

測定の結果，キュー提示付近のタイミングで活動するニューロンが複数存在した．その中に，スケジュール進行度，すなわちキューの明るさに応じた活動を示すものは見られなかったが，キュー（スケジュール進行度）によって異なる応答を示した．応答のタイプには，表 1 に示す 5 種類があった（○は応答があったキューを表す）．ただし，このようなキューに対する応答性はキュー条件だけで見られ，ランダム条件ではキューによる活動の差はなかった．

キュー条件では，スケジュール進行度が大きいほど報酬が「近い」（報酬が得られるまでの試行数が少ない）ので，状態価値が高い．一方，ランダム条件では，報酬はキューに関係なくランダムに与えられるので，状態価値は一定である．このことから，「腹側線条体

表 1 キュー条件における応答（Shidara ら [7] より引用）

Table 1 Response in the cue condition (adapted from Shidara et al. [7]).

	1/3	1/2	2/3	3/3	2/2	1/1	n
(1)			○	○	○		16
(2)	○	○				○	13
(3)	○	○					6
(4)				○	○	○	3
(5)				○	○		3

ニューロンは提示キューそのものではく，提示キューの価値に対して応答している」と解釈されてきた [12].

3. 腹側線条体ニューロン活動の再解析

表 1 において大多数を占める (1)，(2) のタイプに注目すると，前節で述べた解釈とは別の解釈が可能であることに気付く．

タイプ (1) のニューロンが応答を示す（タイプ (2) が示さない）キュー 2/3, 3/3, 2/2 が提示されるのは，必ず，報酬が与えられない試行の後であるのに対して，それ以外のキュー 1/3, 1/2, 1/1 は，ほとんどの場合，前スケジュール報酬が与えられた後に提示される．すなわち，これらのニューロンは，現在のキューではなく,「前試行において報酬があったかどうか」に応答しているという解釈が成り立つ．ランダム条件では前試行の報酬とキューとの相関はないから，ランダム条件においてキュー依存性が失われることとも整合する．

そこでここでは,「過去の履歴」という観点から Shidara らのデータを再解析し，どちらの解釈が妥当か検討する．

3.1 方　　法

キュー条件では多くの場合，過去の履歴から次に提示されるキューを予測できる．そのため，過去の履歴に応答しているのか，現在のキューに応答しているのか，区別が困難である．例えば，**3.2.1** で述べる解析をキュー条件のデータに適用してもほぼ同様の結果が得られるが，履歴ではなく提示されるキューを予測して反応したという可能性を排除できない．そこで，再解析の対象はランダム条件のデータが得られている 26 個のニューロンとし，ランダム条件のデータを中心に解析した．

ニューロンの応答性の指標には応答区間内の発火数を用いる．応答区間にはニューロンによる個体差が存在するため，スパイク密度関数（$\sigma = 10$）に基づき，次のように定義した（図 4）．まず，キュー前 400〜200 ms の区間の平均スパイク密度をオフセットとする（図の点線）．次にキュー前 200 ms〜後 1000 ms の密度関数からオフセットを引く（マイナスの部分は 0 とみなす）．そして，ピーク（細い実線）を中心とした内側 90%の区間をそのニューロンの応答区間（太い実線の間）とする．

3.2 結　　果

3.2.1 応答開始時間

図 5 に，ランダム条件における 26 個のニューロンの応答開始時間のヒストグラムを示す．ここで応答開始時間とは，**3.1** で示した応答区間の最初の時刻（キュー提示時刻を 0 とする）である．

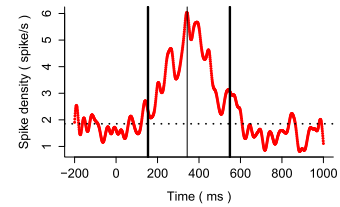


図 4　応 答 区 間
Fig. 4　Response period.

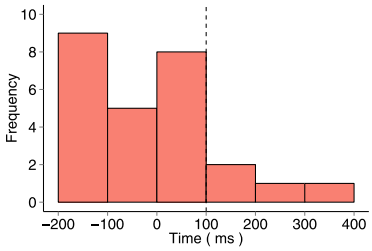


図 5　応答開始時間のヒストグラム
Fig. 5　Histogram of the response onset time.

図からわかるように，過半数（14/26）のニューロンが，キュー提示前から活動を開始している．また，腹側線条体へ視覚情報を伝えると考えられる嗅周皮質における反応潜時のデータ [13], [14] から，キューの情報が腹側線条体ニューロンの活動に反映されるまで 100 ms 以上かかると見積もられるので，応答開始時間が 0〜100 ms である 8 個のニューロンの活動も，提示キューへの直接的な応答とは言えない．提示キューは試行ごとにランダムに選ばれるから，キューを予測することによって提示前から活動を開始したという解釈も成り立たない．したがって，これらのニューロンの活動は，過去の履歴を反映している可能性が高いと考えられる．

3.2.2 履歴依存性

次に，キュー提示以前の情報への依存性を調べるため，応答区間内のスパイク数に関して，キュー提示直前の情報である「前試行報酬」，二つ前の情報である「前試行提示キュー」，三つ前の情報である「2 試行前報酬」の 3 要因による 3 元配置分散分析を行った．

要因「前試行報酬」及び「2 試行前報酬」はそれぞれ「報酬有」と「報酬無」の 2 水準からなる．また，要因「前試行提示キュー」の水準は,「キュー 1」と「キュー

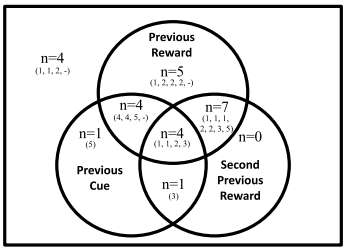


図 6　腹側線条体ニューロンの履歴依存性の分類図
Fig. 6　Classification diagram of history dependence for the ventral striatum neurons.

1 以外（1/2，1/3，2/3）」の二つとした．キューをこの 2 水準に分けたのは，ランダム条件において両者の出現確率が等しいこと，またキュー条件において直後の報酬を予期させる・させないに対応することによる．

解析の結果，26 個中 22 個のニューロンについて何らかの要因に有意差（5%水準）が認められた．そのうち 11 個は主効果のみ（複数の主効果含む）が有意であり，同じく 11 個には交互作用が見られた．

この結果を基に，どの履歴情報に依存性をもつニューロンが何個あったかをベン図に示したのが図 6 である．ここでは交互作用が見られたニューロンは，関係する要因全てに依存性があるとしている（例えば，前報酬と前キューの交互作用が存在するニューロンは，前報酬と前キューの両方に依存性をもつものに分類した）．図中 “n =” の後の数字はニューロンの個数を，その下の括弧内に並ぶ n 個の数字は，それらのニューロンが表 1 の (1)〜(5) のどのタイプに分類されていたものか（“-” はどれにも分類されないもの）を表す．この図から，前試行報酬の有無を中心に，過去のさまざまな情報の組み合わせに依存するニューロンが存在することがわかる．

3.3 考　　察

以上の解析結果に基づいて，腹側線条体で観測された「キュー応答ニューロン」に関する従来の解釈の妥当性を検証しよう．

まず，表 1 のタイプ (1) 及び (2) に分類されたニューロンでランダム条件のデータがあるものは 15 個あったが，そのうち 12 個は前試行報酬への有意な依存性が見られた．また，依存性が有意でなかったものを含めて，タイプ (1) に分類された 8 個は全て前試行報酬がないときの活動の方が大きく，タイプ (2) に分類された 7 個は全て前試行報酬があるときの活動の方が大きかった．キュー条件においてもキュー提示前に活動を開始する場合が多いことも考慮すると，これらのニューロン活動は，主に前試行報酬の有無を反映している可能性が高い．

また，タイプ (3)〜(5) に分類されたニューロンのうちランダム条件のデータがあるものは 8 個であったが，これらは全て前試行キューまたは前々試行報酬への依存性を示した．このことから，これらのニューロンの活動のパターンも，過去の履歴が反映された結果である可能性がある．

ただし，これらは提示キューに対して全く応答しないということではない．応答区間の最初とキュー提示後 100 ms 以降とで活動が変化する場合もしばしばあることから，提示キューに対する応答も一部含まれていると考えられる．

しかしながら，Shidara らの解析において提示キューへの依存性が見られなかったことから，少なくともランダム条件におけるニューロン活動は，過去の履歴を反映していると解釈するのが妥当であろう．腹側線条体において，状態価値が直接表現されているという証拠が少ないという事実と合わせると，腹側線条体（または striosome）の役割を状態価値の表現に限定している点において，従来のモデルには改善の余地があると考えられる．

以上の考察に基づいて，我々は「腹側線条体は，単に現在の入力から状態価値を表現するのではなく，過去の入力の系列から状態価値を推定するのに適した表現に変換する機能をもつ」という仮説を立てた．次章では，この機能のモデル化を行う．

4. 腹側線条体の機能のモデル化

4.1 モデルの構造

腹側線条体ニューロンが過去の刺激（報酬を含む）を反映するということは，履歴情報が何らかの形で保持されているということである．これを実現する最も単純な方法は，過去の入力をバッファに蓄えておく方法であるが，2 試行前の報酬のような過去の情報を保持する回路が脳内にあるという報告はない．また，状態価値の推定にどれだけ古い情報まで必要かは課題によって異なるから，さまざまな課題に対応するためには十分多数のバッファが必要となり，非効率的だと考えられる．

バッファを用いない方法として，リカレント結合を用いて出力を入力側にフィードバックすることが考えられる．これによって常に 1 時刻前の情報を含んだ情

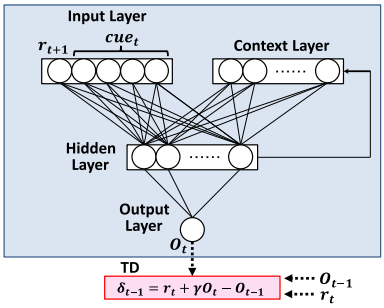


図 7　提案モデルの構造
Fig. 7　Structure of the proposed model.

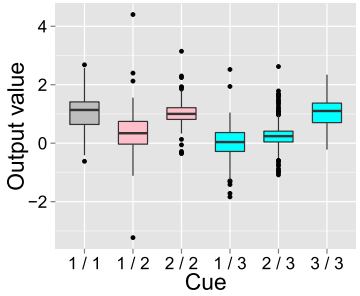


図 8　キュー系列に対するネットワーク出力
Fig. 8　Network output to the test sequence.

報が入力されるため，一度外部から入力された刺激をしばらくの間保持することが可能となる．また，リカレント結合を適切に学習することができれば，入力された刺激の系列から状態価値を推定するのに適した表現を獲得できると考えられる．そこで，ここでは最も単純なリカレントニューラルネットの一つである Elman ネット [15] を用いて，腹側線条体の機能のモデル化を試みる．

モデルの構造を図 7 に示す．ネットワークの入力部は，入力層と文脈層からなる．入力層は直前に受けた外部刺激を表し，文脈層は 1 時刻前の中間層の状態のコピーを保持する．中間層は入力層と文脈層から入力を受け，出力層は中間層からの入力を受けて状態価値を計算する．後ほど詳しく考察するが，この構造は図 1 に示した大脳基底核の回路構造との整合性が高い．

4.2　計算機シミュレーション

4.2.1　方　　法

モデルにキュー条件を模した刺激の系列（学習系列）を入力し，学習させる実験を行った．

入力層には時刻 t におけるキュー cue_t と cue_t 観測後に得られる報酬 r_{t+1} を入力する．また，入力層は五つの素子からなり，それぞれ報酬及び 1，1/2，1/3，2/3 のキューが与えられたときに 1，それ以外のときは 0 を出力する．中間層の素子数は 50 とし，活性化関数にシグモイド関数を用いる．文脈層の素子数は中間層と同じであり，出力層は入力の荷重和をそのまま出力する線形素子 1 個からなる．

通常の Elman ネットでは，正解を教師信号として与えた上でバックプロパゲーション（誤差逆伝播）学習を行うが，強化学習の枠組みでは，正解すなわち正しい状態価値はどこからも与えられない．そこで，ここでは TD 誤差に相当する

$$\delta_{t-1} = r_t + \gamma O_t - O_{t-1} \tag{4}$$

を誤差信号として，これが 0 に近づくようにバックプロパゲーション学習を行う．ここで，O_t は時刻 t におけるネットワーク出力を表し，報酬 r_t は 0 または 1 をとるものとする．なお，脳内での TD 誤差の計算過程がはっきりしないこともあり，今回は単に計算機上で保持した r_t と O_{t-1} 用いて誤差信号を計算した．学習系列の長さは 200，割引率 γ は 0.3 とした．

学習後のネットワークに対して，学習系列とは別の 2 種類の刺激系列を入力して応答を解析する．一つはキュー条件を模したもの（キュー系列），もう一つはランダム条件を模したもの（ランダム系列）であり，いずれも長さは 200 である．

結果の一般性を確保するために，乱数のシードを変更することによって結合荷重の初期値及び学習系列，キュー系列，ランダム系列を変えた上で，実験を 10 回繰り返した．

4.2.2　結　　果

キュー系列に対するネットワークの出力を図 8 に示す．全 10 回の実験に関して，各キューが入力されているときの出力の中央値とばらつきを表している．この図から，1/2 → 2/2，1/3 → 2/3 → 3/3 のように，報酬に近い状態ほど出力値が増加していることがわかる．このことから，ネットワークの出力が状態価値を表すように学習がなされたと言える．

一方，ランダム系列を入力した場合には，キューによる出力値に有意な違いは見られなかった．そこで中間素子の出力について，**3.2.2** と同様に「前試行報酬」,「前試行提示キュー」,「2 試行前報酬」の 3 要因で 3 元配置分散分析を行ったところ，ほとんどの素子が履歴に依存した出力の変化を示すことがわかった．図 6 と同様の分類を行った結果を図 9 に示す（図中

の数値は各カテゴリーに分類された素子数の 10 回の実験の平均値）．腹側線条体ニューロンの場合と同様に，さまざまな履歴の組合せに依存する素子が広く分布していることがわかる．

図 10 は，あるランダム系列を入力したときの中間素子の応答例であり，入力系列に対する素子の出力値を 4 グループに分けてプロットしたものである．図 10 (a) は前試行報酬及び 2 試行前報酬の主効果が有意であった素子（前試行報酬，$F(1,190) = 34.1$，$p < 0.01$; 2 試行前報酬，$F(1,190) = 19.1$，$p < 0.01$）であり，左から順に 2 試行前報酬と前試行報酬がそれぞれ「無・無」,「無・有」,「有・無」,「有・有」を表す．2 試行前報酬，前試行報酬共にない場合に出力値が大きいことがわかる．また，図 10 (b) は前試行報酬と前試行キューの交互作用が見られた素子（$F(1,190) = 4.99$，$p < 0.05$）であり,「前キューが 1」のとき（左パネル）には前試行報酬の影響は見られないが,「前キューが 1 以外」のとき（右パネル）には前試行報酬の有無が出力に影響を及ぼしている．

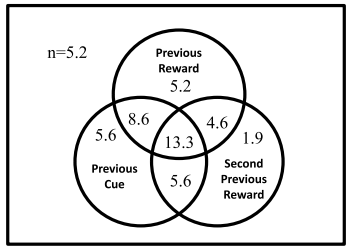
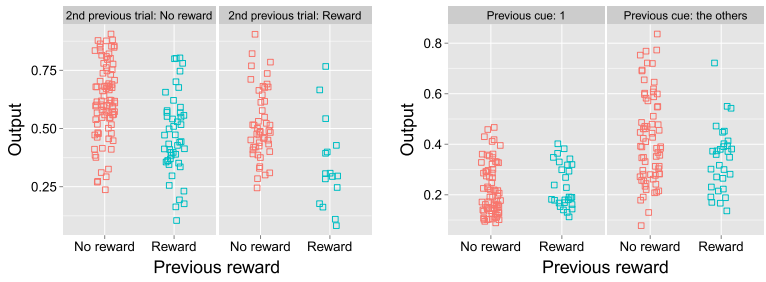


図 9　モデルの中間素子の履歴依存性の分類図
Fig. 9　Classification diagram of history dependence for the middle elements of the model.

こうした応答によく似た応答を示すニューロンが実際の腹側線条体にも見られる．図 10 に示した素子と分散分析の結果が同じであったニューロンの例を図 11 に示す（(a): 前試行報酬，$F(1,145) = 15.9$，$p < 0.01$; 2 試行前報酬，$F(1,145) = 4.21$，$p < 0.05$, (b): $F(1,227) = 4.36$，$p < 0.05$）が，**3. 2. 2** で解析した腹側線条体ニューロンのほとんどについて，類似した応答の素子をモデルに見出すことができた．

4. 3　考　　察

以上の結果から，提案したモデルは，想定した腹側線条体の機能を実現するとともに，ランダム条件におけるニューロン活動の履歴依存性を再現できることが



(a)　(b)

図 10　ランダム系列に対するモデルの中間素子の応答例
Fig. 10　Example of the response of middle elements to a random sequence.

(a)　(b)

図 11　ランダム条件における腹側線条体ニューロンの応答例
Fig. 11　Example of the response of ventral striatum neurons in the random condition.

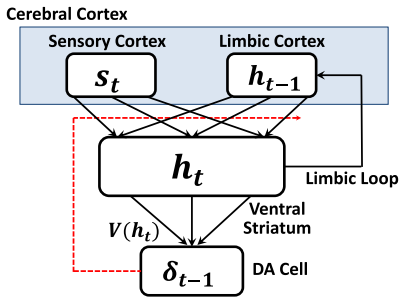


図 12　提案モデルと脳内構造の対応
Fig. 12　Correspondence of the proposed model to the brain structure.

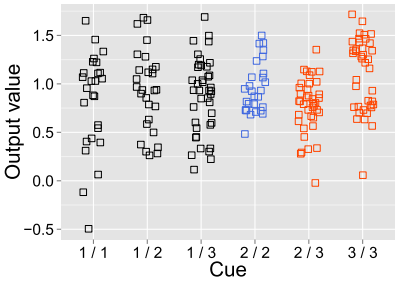


図 13　内部状態から推定した状態価値
Fig. 13　State values estimated from the internal state.

わかった．

このモデルは，腹側線条体がもつと考えられる機能の一つをモデル化したものであって，大脳基底核のモデルとは必ずしも言えない．しかし，関連する生理学及び解剖学的知見とかなりの整合性がある．

まず，図 12 に示すように，提案モデルを大脳基底核の構造と対応づけることが可能である．モデルの要点であるリカレント回路は，辺縁系を介した神経投射がその役割を果たすのではないかと考えている．モデルの中間層に対応する腹側線条体の出力がそのまま辺縁系皮質にコピーされるとは考えにくいが，トポグラフィックな結合が形成されていることから，大部分の情報が保持されたままフィードバックされる可能性は十分にある．

また，提案モデルは，外部からの入力が無くても内部状態からある程度状態価値を推定することが可能である．図 13 に，学習後のネットワークに状態価値を推定させた結果を示す．これは，あるキュー系列の刺激を順に入力する際，一時的に入力層の値を全て中立値にして，文脈層の信号のみから計算した出力値の分布を示しており，横軸は直後に入力されるキューを表す．

分散分析の結果，キューの違いによる出力値の差は有意であった（$F(5, 193) = 3.53$，$p < 0.01$）．また，キューが 2/2 及び 3/3 のときに相対的に高い値を示し，平均値はキューが 2/3 の場合よりも有意（3/3 vs 2/3，$t(70) = 4.41$，$p < 0.01$，2/2 vs 2/3，$t(60) = 2.6$，$p < 0.01$）に大きかった．ただし，1/1, 1/2, 1/3 のキューはランダムに選ばれるため，その直前の予測はもともと不可能であるし，その他の場合も出力値のばらつきが大きく，必ずしも状態価値を正しく推定できているとは言えない．その理由の一つとして，入力を切った状態での学習を行っていないことが考えられる．しかし，これは実際のニューロン活動の場合も同様であり，キュー提示前の活動が状態価値を正しく反映しているとは限らない（キュー提示の前後でしばしば活動が大きく変化する）し，サルにとってキュー提示前に正確な予測が必要な課題ではない．いずれにせよ，本モデルの素子が刺激を入力しなくてもキュー依存的な活動を示すということは，腹側線条体の「キュー応答ニューロン」が，キュー提示前から活動しうる理由を説明する．

更に，学習後のモデルの中間層は，状態価値の推定に適した表現となるため，素子の中には状態価値に比較的近い応答をするものも一部存在する．したがって，腹側線条体において将来の報酬を反映した活動が見られたとする過去の生理学的知見 [10], [11] も説明可能である．

一方で，本モデルには幾つかの限界がある．まず，本研究の直接の目的ではないものの，キュー条件におけるニューロン活動の再現は本モデルではされなかった．すなわち，キュー系列を入力したときの中間素子の活動は，ランダム系列を入力したときと同様に，多様な活動パターンを示し，表 1 に示す幾つかの活動パターンに集約されることはなかった．逆に，表 1 のような活動パターンを再現するだけであれば，原理上，本モデルのような回帰結合は不要である．したがって，ランダム条件に加えてキュー条件におけるニューロン活動も再現するためには，回帰結合を受けない素子を追加する，キュー条件では直接入力の強度を高める，といったモデルの修正が必要かもしれない．

また，大脳基底核のモデルとして不十分な点が幾つかある．一つは，状態価値 V の表現場所である．モデルでは，出力層の素子の出力が V を表しているが，図 12 のように対応づけると，腹側線条体からドーパ

ミンニューロンまでの途中で表現されていることになる．これは,「腹側線条体では，価値関数の基底関数が表現されている」という鮫島ら [12] の仮説に似ているが，生理学的な証拠はまだ得られていない．また，行動選択など TD 誤差の計算以外に V を使うのが難しいという問題もある．したがって，状態価値 V あるいは行動価値 Q が，辺縁系皮質や線条体の別の領域（matrix など）に別途表現されると考える必要があるかもしれない．

もう一つの問題点は，モデルではバックプロパゲーション学習を行っていることである．一般的なバックプロパゲーション学習は生物学的妥当性が乏しく，それがそのまま脳内で行われているとは考え難い．しかし，提案モデルの場合，出力層の素子が一つであり，誤差自体はドーパミン投射の形でフィードバックされるので，問題となるのは中間層から出力層への結合荷重が学習によって変化し，その値が中間素子の学習に必要だという点だけである．これについては，中間層から出力層の結合を固定値とする代わりに有効な中間素子の数が変わるようにする，といった方法によって，無理のない形で同等な学習が実現できるのではないかと考えている．

最後に，提案モデルの計算論的な意味について考察する．従来のモデルでは，大脳皮質からの入力を現在状態とし，そこから直接状態価値を求めていた．しかし，大脳皮質が受ける刺激入力は，常に変化し，全く同一の刺激が入力されることはない．したがって，ある特定の刺激入力に注目するなどの方法によって「同じ状態」が大脳皮質で認識されることを前提としている．しかし，どの刺激が報酬予測に重要なのか，サルが事前に知っているわけではない．例えばキュー条件において提示されるキューの種類が「状態」に対応するというのは，課題を設定した人間の考えにすぎない．ランダム条件でのニューロン活動を見ると，サルは常に報酬を予測する手がかりを探っているように思われる．

これに対して，提案モデルでは，どの刺激が報酬予測にどの程度重要か，わからないことを前提としている．その上で，報酬を予測すると同時に，中間層に状態価値推定に適した状態空間を構成するものとみなすことができる．実際，このモデルと同様なリカレント型ニューラルネットを用いて，状態空間を構成しつつ強化学習を行うモデルが提案されており [16]，未知環境における行動学習に有効であることが示されている．同様な機能が大脳基底核にあるならば，神経科学的に興味深いだけでなく，工学的な意義も大きいと言えよう．

5. むすび

サルの腹側線条体の神経活動データを再解析し，腹側線条体ニューロンが，予測される報酬だけではなく,「前試行報酬」,「前試行キュー」,「2 試行前報酬」といった過去の履歴の組み合わせに依存した応答を示すことを明らかにした．この結果から，腹側線条体は過去の履歴から状態価値を推定するための中間表現を保持する，という仮説を立て，そのような機能をリカレント型ニューラルネットによってモデル化した．計算機シミュレーションの結果，構築したモデルは過去の入力から将来の報酬を予測できるだけでなく，腹側線条体の神経活動パターンを再現することがわかった．

神経構造との対応などから，大脳基底核において提案モデルと同様な機能が実現されている可能性は十分にある．また，腹側線条体のニューロン活動に関する種々の知見に対して，統一的な説明を与える．このことは，大脳基底核において，刺激入力の時系列から状態価値の推定に適した状態空間を構成しながら強化学習が行われている可能性を示唆するが，このような視点は従来のモデルになかったものである．

今後の課題として，まず **4. 3** で述べたモデルの問題点を解消することが挙げられる．また，多試行報酬スケジュール課題以外の実験課題についてシミュレーションを行い，生理データと比較することも検討している．そのほか，本研究の結果に基づいて，従来の強化学習モデルの詳細を再検討することも重要な課題である．例えば，従来 TD 誤差を計算するために，1 時刻前の状態価値を保持するバッファを必要としていたが，提案モデルでは 1 時刻前の内部状態が辺縁系皮質から入力されるため，これを用いてより自然な方法で TD 誤差が計算できるかもしれない．

大脳基底核の強化学習モデルには，生理学的な裏付けが十分でない部分や，計算論的に見て不十分と思われる部分がまだ多くある．一方で，大脳基底核に関する生理学的知見の中には，計算論的な検討が十分になされてないものも多い．本研究で行ったように，新たな観点で生理データを見直し，その結果に基づいてモデルを修正することが今後重要だと思われる．

謝辞　本研究の一部は，科学研究費補助金特定領域研究（課題番号 17022052）及び基盤研究 (B)（22300079,

22300138, 25282246）の支援を受けて行われた．

文　献

[1] R.S. Sutton and A.G. Barto, Reinforcement Learning, MIT Press, 1998.

[2] W. Schultz, P. Dayan, and P.R. Montague, “A neural substrate of prediction and reward,” Science, vol.275, pp.1593–1599, 1997

[3] A.G. Barto, “Adaptive critics and the basal ganglia,” in Models of Information Processing in the Basal Ganglia, ed. J.C. Houk, J.L. Davis, and D.G. Beiser, pp.215–232, MIT Press, 1995.

[4] K. Doya, “Complementary roles of basal ganglia and cerebellum in learning and motor control,” Current Opinion in Neurobiology, vol.10, no.6, pp.732–739, 2000.

[5] B.L. Goldstein, B.R. Barnett, G. Vasquez, S.C. Tobia, V. Kashtelyan, A.C. Burton, D.W. Bryden, and M.R. Roesch, “Ventral striatum encodes past and predicted value independent of motor contingencies,” Journal of Neuroscience, vol.32, pp.2027–2036, 2012.

[6] Y.B. Kim, N. Huh, H. Lee, E.H. Baeg, D. Lee, and M.W. Jung, “Encoding of action history in the rat ventral striatum,” J. Neurophysiology, vol.98, pp.3548–3556, 2007.

[7] M. Shidara, T.G. Aiger, and B.J. Richmond, “Neuronal signals in the monkey ventral striatum related to progress through a predictable series of trials,” J. Neuroscience, vol.18, pp.2613–2625, 1998.

[8] C.R. Gerfen, “The neostriatal mosaic: Multiple levels of compartmemtal organization in the basal ganglia,” Annual Review of Neuroscience, vol.15, pp.285–320, 1992.

[9] J.N.J. Reynolds, B.I. Hyland, and J.R. Wickens, “A cellular mechanism of reward-related learning,” Nature, vol.413, pp.67–70, 2001.

[10] W. Schultz, P. Apicella, E. Scarnati, and T. Ljungberg, “Neuronal activity in monkey ventral striatum related to the expectation of reward,” Journal of Neuroscience, vol.12, pp.4595–4610, 1992.

[11] H.C. Cromwell and W. Schultz, “Effects of expectations for different reward magnitudes on neuronal activity in primate striatum,” J. Neurophysiology, vol.89, pp.2823–2838, 2003.

[12] 鮫島和行，銅谷賢治，“強化学習と大脳基底核，” バイオメカニズム学会誌，vol.25, no.4, pp.167–171, 2001.

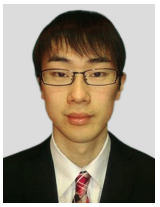
[13] Z. Liu and B.J. Richmond, “Response differences in monkey TE and perirhinal cortex: Stimulus association related to reward schedules,” J. Neurophysiology, vol.83, pp.1677–1692, 2000.

[14] Y. Naya, M. Yoshida, and Y. Miyashita, “Forward processing of long-term associative memory in monkey inferotemporal cortex,” J. Neuroscience, vol.23, pp.2861–2871, 2003.

[15] J.L. Elman, “Finding structure in time,” Cognitive Science, vol.14, pp.179–211, 1990.

[16] Y. Sawatsubashi, M.F.B. Samusudin, and K. Shibata, “Emergence of discrete and abstract state representation in continuous input task through reinforcement learning,” Advances in Intelligent Systems and Computing, vol.208, pp.13–22, 2013.

（平成 26 年 11 月 11 日受付，27 年 3 月 17 日再受付，
6 月 2 日早期公開）



篠塚　正成

平 26 筑波大学大学院システム情報工学研究科博士前期課程修了．在学中，脳の情報処理機構の研究に従事．

森田　昌彦　（正員）

昭 61 東大・工・計数卒．平 3 同大大学院博士課程修了．日本学術振興会特別研究員，東京大学工学部助手を経て，平 4 筑波大学電子・情報工学系講師．同大機能工学系助教授などを経て，平 19 同大大学院システム情報工学研究科教授．現在，同大システム情報系知能機能工学域に所属．脳の情報処理機構及び神経回路網による情報処理の研究に従事．平 5 日本神経回路学会研究賞，平 6 同学会論文賞，平 11 日本心理学会研究奨励賞受賞．

設楽　宗孝

昭 59 東大理学部生物学科卒．昭 61 東大大学院理学系研究科動物学専門課程（修士）修了．平成 2 東大大学院医学系研究科（博士）修了，医学博士．平 2 電子技術総合研究所，平 13（独）産業技術総合研究所を経て，平 17 より筑波大学大学院人間総合科学研究科教授．現在，同大医学医療系生命医科学域に所属．報酬系と行動決定，及び視覚認識のシステム脳科学研究に従事．平 6 日本神経回路学会論文賞受賞．