動的環境における進化と学習の相互作用
～Baldwin 効果の工学的応用に向けて～

鈴木 麗璽
名古屋大学 大学院人間情報学研究科
reiji@info.human.nagoya-u.ac.jp

Abstract

進化と学習の相互作用に関する重要なトピックに，Baldwin 効果と呼ばれる現象がある．我々は，個体間の相互作用のみを考慮した動的な環境として繰り返し囚人のジレンマゲームにおける戦略の進化を取り上げ，戦略に表現型の可塑性を導入し進化実験を行っている．本稿では，このような動的環境においても進化の過程において Baldwin 効果が有効に働き，協調関係を維持するために適度な可塑性を持った安定した集団へと進化したことを示す．また，得られた知見を踏まえて Baldwin 効果の工学的応用について検討する．

1. はじめに

Baldwin 効果は，およそ 100 年ほど前，Lamarck 的な獲得形質の遺伝の仕組みを用いず，自然選択のみによって進化と学習の相互作用を示すものとして Baldwin によって提案された[1]．これは，進化と学習が相互に与える影響を学習のメリットとコストのバランスから説明するものであり，現在の一般的な定義では，次の 2 つの段階を経て，学習により獲得されていた形質が次第に生得的な形質へと進化していくものとされている[2].

なお，ここで述べる学習とは，人間の知能に代表されるような高度な学習メカニズムだけでなく，たとえば運動による筋肉の増強や日焼けによる皮膚の色の変化などを含む表現型の生涯の変化，すなわち表現型の可塑性を示すものである．

- 第 1 段階：学習により生存上有利な形質を獲得した個体が次世代に多く子孫を残す.
- 第 2 段階：十分多くの個体が生存上有利な形質を学習により獲得した集団では，学習にかかるコストのためその形質を生得的に獲得している個体が次世代に多く子孫を残す.

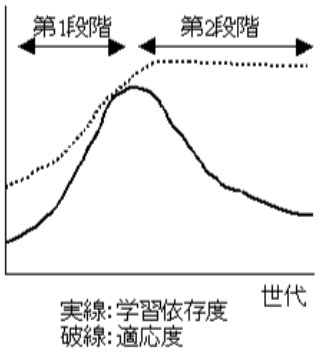

図 1：Baldwin 効果

第 1 段階は，学習によるメリットが中心的な選択圧として働いた状態，第 2 段階はコストが選択圧として働いた状態であり，この結果進化のスピードを加速する可能性があるとされている．このとき，集団全体の学習に対する依存度というものが定義できるとすれば，2 つの段階を経て，典型的には図 1 のような適応度と依存度のカーブを描くと考えられる．近年，Hinton と Nowlan による先駆的な進化実験[3]によりこの効果が明確にされて以来，この効果は生物学的側面からだけでなく，進化的計算などの工学的分野からも注目されるようになり，新たな局面を迎えている．

これまで，Baldwin 効果に関する進化実験を用いた研究は，彼らのモデルをはじめとして最適解が固定されたものや，進化と学習のアナロジーを遺伝的アルゴリズムにおける全域探索と局所探索として捉え，学習が進化の過程に対して補助的な役割を果たすというような設定が多かった．また，モデルにおいて学習のメリットやコストを明示的に導入したものがほとんどであった．しかし，現実世界は動的な環境であり，学習は刻々と変化する状況に対して柔軟に振る舞う術を与えてくれる．また，学習は常に良い結果ばかりをもたらすわけではなく，環境の変動が誤った学習や不必要な学習を引き起こす状況も考えられる．

そこで本研究では，学習によって得られるメリット（及びコスト）が世代を通して保証されないような動的な環境においても，上記のような進化と学習の相互作用を Baldwin 効果として積極的に捉え，動的な環境における進化と学習の相互作用について知見を得ることを目的とする[4]．また，今回は特に個体間の相互作用に依存し

て各個体の適応度が決定され，世代を通して最適な解が決定できないような設定を採用する．このような環境における進化と学習の相互作用に関する解析結果は，近年注目されているマルチエージェント環境による協調作業システムの進化的獲得などへの応用も期待できると考える．

2. 表現型の可塑性を導入した繰り返し囚人のジレンマゲームの戦略の進化モデル

以上を踏まえ，世代を通して最適解が決まらず，特に個体間の相互作用のみを考慮した動的環境として，遺伝的アルゴリズムを用いた繰り返し囚人のジレンマゲームの戦略の進化モデルを構築した．繰り返し囚人のジレンマゲームついて簡単に説明した後，モデルについて解説する．

2.1 繰り返し囚人のジレンマゲーム

繰り返し囚人のジレンマゲームは，2 人非ゼロ和ゲームの一種で，Axelrod による研究[5]をはじめとして利己的集団における協調行動の創発に関して数多くの研究がなされている．ゲームは表 1 に代表される利得行列を用いて以下の手順で行われる．

- 2 人のプレイヤーは協調（Cooperate）または裏切り（Defect）のどちらかの手を同時に出す．
- 出した手に応じて，利得行列（表 1）から両者が得る得点が決まる．
- この対戦を繰り返し行い，その合計（平均）得点を競う．

表 1：囚人のジレンマゲームの利得行列

相手の手（→） 自分の手（↓）	協調 (C)	裏切り (D)
協調（C）	(R=3，R=3)	(S=0，T=5)
裏切り（D）	(T=5，S=0)	(P=1,P=1)

2R>T+S

（自分の得点，相手の得点）

1 回きりの対戦において，プレイヤーがともに自らの期待利得を最大にするような戦略，つまりこのゲームでの支配戦略である裏切り（D）を取った場合，裏切り合いとなりこれはナッシュ均衡解である．にもかかわらず，この解はパレート最適ではなく，双方にとってより良い解すなわち協調し合いが存在するため，裏切りは正しい判断ではなかったのではないかというジレンマが生じる．さらに，十分長い繰り返しゲームにおいては，交互に裏切るよりも協調し合ったほうが双方の利益となる（2R>T+S）ため，いかにして協調関係を築くことができるかが高い得点を得る際の問題となるが，協調関係を築くことができるかどうかは相手の出方次第である．つまり，ゲームにおいてある戦略がうまくやれるかどうかは，対戦相手に大きく依存する．したがって，ジレンマゲームの戦略集団における総当たり戦の得点を適応度とするような環境を考えると，それは各個体の適応度が世代ごとに刻々と変化する動的な環境として捉えることができる．

2.2 戦略の遺伝子表現

集団における各エージェントは繰り返し囚人のジレンマゲームの戦略を遺伝子として持つ．このモデルでは，各個体の持つ戦略を戦略をあらわす遺伝子列 GS と可塑性をあらわす遺伝子列 GP の 2 つの遺伝子列の組で表現する．戦略をあらわす遺伝子列は Lindgren のモデル[6]と同様な，履歴に依存して次回の手を決定する戦略を定義する．記憶長 m の戦略は裏切りを 0，協調を 1 として以下のような 2 進数で表された履歴 h_m を持つ．

$$h_m = (a_{m-1}, \cdots, a_1, a_0)_2 \qquad (2)$$

ここで a_0 は前回の相手の手，a_1 は前回の自分の手，a_2 は前々回の相手の手…とする．
ある履歴 k に対応して次回出すべき手を A_k（0 または 1）とすると，記憶長 m の戦略は，

$$GS = [A_0 A_1 \cdots A_{n-1}] \quad (n = 2^m) \qquad (3)$$

と表すことができる．これを GS とする．さらに，各 A_x に対してその表現型（協調または裏切り）が可塑性を持つかどうかを P_x（0：可塑性を持たない，1：可塑性を持つ）として，GP を

$$GP = [P_0 P_1 \cdots P_{n-1}] \tag{4}$$

と定義する．例えば，しっぺ返し戦略（初回は協調，以降は前回相手が出した手を真似る）を記憶長 2 で表すと，GS=[0101]，GP=[0000]となる．

2.3 メタ・パブロフ学習

可塑性を持つ表現型は，対戦中にその表現型を用いた結果に応じて，学習により表現型を変更する．ここで，以下のような学習行列（表２）を定義し，この行列に基づいて表現型を変更するという学習を導入する．この行列により，戦略は過去の履歴に依存してパブロフ戦略[6]的に学習を行うためこの学習をメタ・パブロフ学習と呼ぶ．学習は次の手順で行われる．

- 繰り返し対戦を行う前は，各個体は GS の表す純粋戦略をそのまま表現型として持つ．
- 表現型と履歴を参照し対戦を行い，用いた表現型（C または D）に対応する可塑性をあらわす遺伝子列のビットが“１”（可塑的）であった場合，その表現型を対戦結果に対応するメタ・パブロフ学習行列の値（C または D）と置き換えたものを新たな表現型とする．
- 次回の対戦以降，新たな表現型を参照し，手を決定する．

この行列（の値）は，プレイした結果得られる得点が相対的に高ければそのまま変更せず，逆に小さければ変更するという強化学習の原理に基づくものであり，学習則としてシンプルかつ典型的なものとして今回採用する．この行列自体はパブロフ戦略（初回は協調，以降は対戦結果が相対的に良ければ次回も同じ手を出し，悪ければ手を変える）[7]と同じであるが，直前の対戦結果に応じて次回出す手を決定するのではなく，表現型を用いた結果に応じて戦略自体を変更（学習）するという意味で，この行列を用いた学習方式をメタ・パブロフ学習と呼ぶ．

ここで，メタ・パブロフ学習の例として，GS=[0001]，GP=[0011]の戦略が学習する例を示す．図２（学習前）はこの戦略の表現型を図示したものである．記憶長 2 の履歴（前回の自分の手と相手の手）に対応して，次回出す手が表現型として決められている．ただし可塑性を持った表現型には下線を引いた上で，初期状態を示している．過去の対戦履歴が CC であったと仮定すると，表現型からこの戦略は C を出す．このとき相手が D を出したと仮定する．ここで，C を出すのに用いた表現型は可塑性を持つのでメタ・パブロフ学習行列をもとに表現型を変更する．この場合，自分の手が C，相手の手が D なので，学習行列から表現型を D に変更し，次回対戦履歴が CC の場合には D を出すようになる．従って，戦略の表現型は図２（学習後）のように変化する．このように，可塑性をあらわす遺伝子列に 1 のビットを持つ戦略個体は，繰り返し対戦を通して表現型が変化しうるという意味で可塑的な戦略であると捉える．

対応する GP のビットが 1 である GS の値は表現型の初期値としてのみ働く．そこで，今後各戦略を，可塑性を持つ表現型に対応する戦略遺伝子を x と置き換えた遺伝子列でまとめて表現することで，進化の過程の大枠を捉えることにする（e.g. GS＝[1000]，GP＝[1001]→[x00x]）．

なお，本研究では，学習すること自体にかかる明示的なコストを導入せず，学習にかかるコスト（及びメリット）はすべて個体間の相互作用の結果として与えられるものと考える．

表２：メタ・パブロフ学習行列

相手の手（→） 自分の手（↓）	C	D
C	C	D
D	D	C

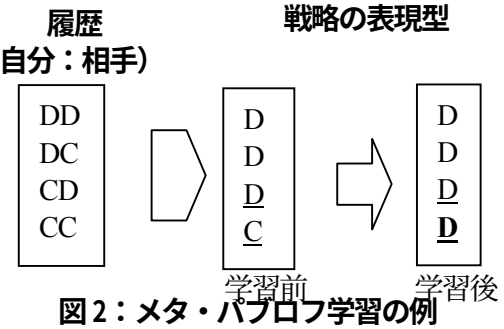

図２：メタ・パブロフ学習の例

2.4 繰り返し対戦と進化

以上のような戦略個体同士でノイズありの繰り返し対戦を表 1 の利得行列を用いて行う．ノイズとは，繰り返し対戦において，各戦略個体が出すべき手が一定の確率で反転してしまうことで，現実世界における表現の間違い，転送経路のノイズ，誤解などの不可抗力を象徴するものである．前節で示したとおり，本研究で用いられる戦略が手を決定するためには履歴が必要である．そこで，各繰り返し対戦の一番初めは，繰り返し対戦ごとにランダムに作成された仮想の履歴を各個体が参照し，初回の手を決定するものとする．繰り返しゲームを行う状況として，「十分長い間繰り返されるが，実際何回繰り返して行われるかはプレイヤーには分からない」という設定にするため，繰り返しの回数は固定せず，対戦ごとに一定の確率で次回の対戦が行われるものとする．この確率を未来係数と呼ぶ．また，可塑的な戦略における表現型は，繰り返し対戦ごとに初期状態（GS が示すままの状態）に戻されるものとする．

このような繰り返し対戦を集団全体において総当たりで行い，その合計得点を各戦略個体の適応度とする．最後に，各適応度に応じたルーレット選択により次世代の集団を生成する．その際，一定の確率で遺伝子のビットが反転する，一点突然変異を導入する．なお，計算量を軽減するために，はじめて行う対戦カードの場合は，繰り返し対戦を 20 回行った平均得点を用いるとともに保存し，すでに行ったことのある対戦カードでは保存した得点を利用するものとする．また，保存した得点は 500 世代ごと消去し，新たに計算し直すものとする．

3．実験結果と考察

記憶長 2（初期集団はランダム）の集団において，パラメータとして突然変異率 1/1500，個体数 1000，ノイズ率 1/25 未来係数 99/100，世代数 2000 を用いて進化実験を行った．その結果，試行の約 70%で図 3, 4 のような傾向を持つ進化が確認された．なお，集団中を占める戦略個体は，可塑的な表現型に対応する戦略遺伝子を x と置き換えた GS でまとめて表現してある（e.g. GS＝[1000], GP＝[1001]→[x00x]）．

はじめの約 60 世代では，裏切り的な戦略が集団中を占め，平均得点（白実線）が低下した．その後，集団の可塑性（GP 中に占める”1”のビットの割合で，このモデルにおける学習依存度を表す指標，黒実線）の増加とともに平均得点は上昇し，約 250 世代までに高い平均得点を維持する協調的な集団へと進化した．これは，可塑性が裏切り的な集団から協調的な集団へのシフトに有利な方向へと働いたことを示し，Baldwin 効果の第 1 段階と考えられる．

その後，高い平均得点を維持したまま，集団の可塑性は次第に低下し約 50%のところで安定し，集団の大部分を[x00x]型の個体が占める結果となった．これは，十分協調関係が築かれた集団においては，ノイズによって可塑性がコストとして働くため，協調集団を維持するために最低限必要な可塑性のみが選択されたことを示しており，Baldwin 効果の第 2 段階と考えられる．

集団の可塑性と平均得点の相関に注目して進化の過程を観察すると，Baldwin 効果の 2 つの段階をはっきりと把握することができる．図 5 は 10 回の試行における，集団の可塑性と平均得点の相関の軌跡を表したものである．初期集団の状態から，一旦平均得点と集団の可塑性が低い状態へと進化した後，共に上昇する方向（グラフ右上）へと進化しているのがわかる．これが Baldwin 効果の第 1 段階である．その後，軌跡はまっすぐ左へと向きを変え，平均得点を維持したまま，集団の可塑性のみが低下しているのがわかる．これが第 2 段階である．

この相関図の軌跡にあわせて，出現した戦略を大まかに分類すると，図 6 のようになる．はじめに，初期集団においては裏切り的な戦略に対して搾取される戦略が多く含まれるため，相関図左下のような全面裏切り的な戦

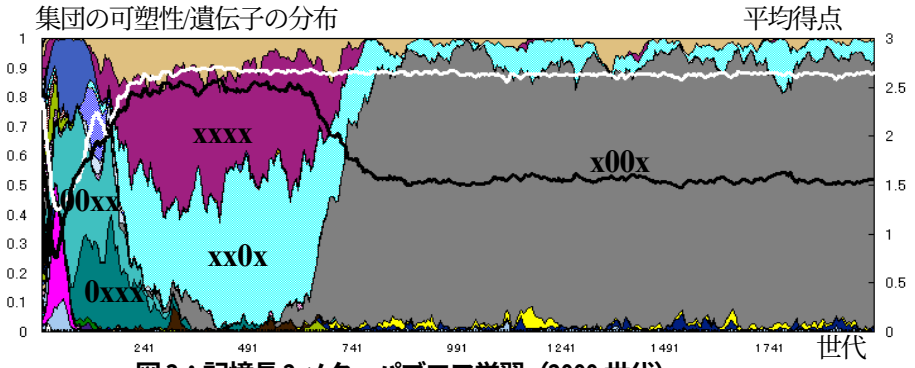

図 3：記憶長 2 メタ・パブロフ学習（2000 世代）

略が有利となり，集団中に広まる．左下の裏切り的な戦略群を中心とした戦略が集団中を占めると，対戦が裏切り合いばかりになり平均得点は低下する．この状態では，右下のような戦略が徐々に集団中に広まる．これらの戦略は同種同士の対戦では戦略に含まれる可塑性によってノイズや初期状態をきっかけにして協調を出し合うので，裏切り合いと比べて若干高い得点を得るためである．その後，図右下のような戦略が増加し集団全体においてさらに協調する機会が増え，平均得点が上昇してくると，裏切り的な戦略に対してそれほど点を与えないことを維持しつつも，右下の戦略と比べてより協調的であることで高い得点を得る右上のような，より可塑的な戦略が集団中を占めるようになる．

ここで，最終的に集団の大半を占める[x00x]型戦略よりも右上の戦略群の方へと進化する傾向が高いのは興味深い．左下から右下，右上の戦略群への進化のように平均得点が上昇している状況では，他種の得点を下げるよりもまず自身の得る得点が高いことが集団中に広まるために必要とされる．しかし，[x00x]の ESS 的な性質[4]が右下の戦略群との対戦で両者の得る得点を，右上の戦略群と比べて下げてしまっているので，右上の戦略が先に集団中を占めると考えられる．このことは，ESS 的な戦略であるからといって，他の戦略集団に容易に入り込むことができる訳ではないことを示している．これまでの可塑的な協調集団に至る過程が Baldwin 効果の第 1 段階である．

その後，右上のような協調的な戦略ばかりになると，対戦のほとんどが協調し合いになり，異なる戦略同士の適応度の差が小さくなる．この状態においてほとんどの対戦は，基本的に協調しあい，ノイズが入ると一定の回復過程を経て協調しあいにもどるというサイクルになるため，戦略の差はノイズが入ってからの振る舞いに現れる．このとき，ノイズをきっかけに自分にとって不利な協調を出す可能性のある余分な可塑性，すなわちコストとして働く可塑性が取り除かれていき，最終的には「協調集団の維持に必要最小限の可塑性を持った戦略」[x00x]型戦略が徐々に集団中を占める．[x00x]は記憶長２の戦略群においては ESS 条件を満たすため，ここで集団は安定する．この過剰な可塑性の減少と安定化が，Baldwin 効果の第 2 段階である．

以上から，この過程において，学習は裏切り的な戦略集団から協調的な戦略集団へのシフトと安定化に大きな影響を与えていると考えられる．また，戦略における可塑性は，対戦する戦略に応じて自身の振る舞いをうまく変える特徴を実現するのに有効であると考えられる．

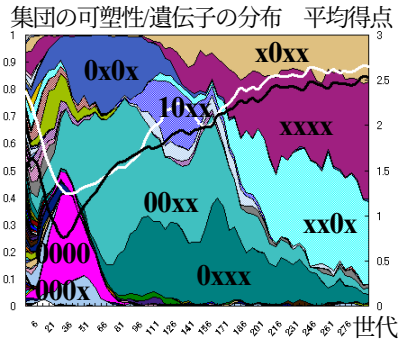

図 4：記憶長 2 メタ・パブロフ学習（300 世代）

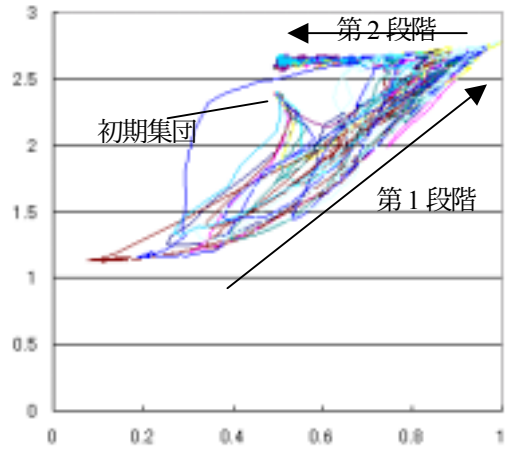

図 5：進化の過程における集団の可塑性と平均得点の相関

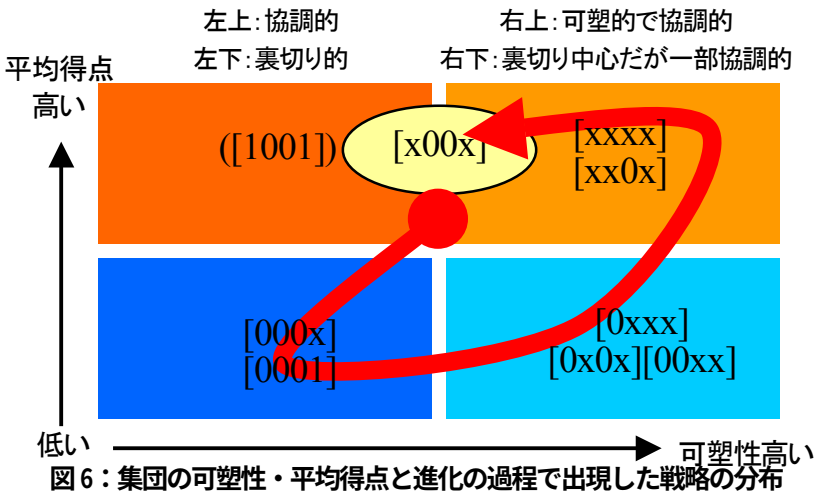

図 6：集団の可塑性・平均得点と進化の過程で出現した戦略の分布

４．より一般化した進化実験 ～学習行列を遺伝子に取り込んだ進化～

これまでの学習行列を用いた実験では，学習行列の値としてメタ・パブロフ学習行列を定義し，前もって与えてきた．しかし，これ以外の学習行列を持つ戦略がどのように振る舞うかは興味深い点である．そこで学習行列を表す遺伝子列 GT を新たに 3 つ目の遺伝子列として定義し，各個体が保持するものとして進化実験を行った．GT は表 3 のように定義される（例：メタ・パブロフ学習行列:GT=[1001]）．

表 3：学習行列をあらわす遺伝子列 GT の定義

相手の手（→） 自分の手（↓）	C	D
C	(CC)	(CD)
D	(DC)	(DD)

GT=[(DD)(DC)(CD)(CC)]

以上の変更をモデルに加え，初期集団は GS,GP,GT 各遺伝子の値をランダムに決定した 100 種をそれぞれ 10 個体ずつ用意するものとして進化実験を 4000 世代に渡って行った(図 6). 図中の####:****はこれまでの遺伝子表記における記述が####，GT が****の個体であることを示す．

60 回試行を行ったところ，そのうち 29 回で 4000 世代の時点で[x00x] GT=[1001]すなわちメタ・パブロフ学習行列を用いた戦略が集団中を占めた（図 7）．また，そのうち 3 回で，Boerlijst らが提案した pPAVLOV 戦略[8]とほぼ同等の振る舞いをする[x001]GT=[1000]が集団中を占めた（図 8）．残りの 28 回ではひとつの戦略には収束しない不安定な状態であった．

学習行列を限定しなくても最終的にメタ・パブロフ[x00x]型戦略が集団中を占めたことから，メタ・パブロフ学習行列は学習則として妥当なもののひとつであると言える．また，このようなシンプルなモデリングにおいて，pPAVLOV のような手作業で作られた戦略が現れたのは興味深い．

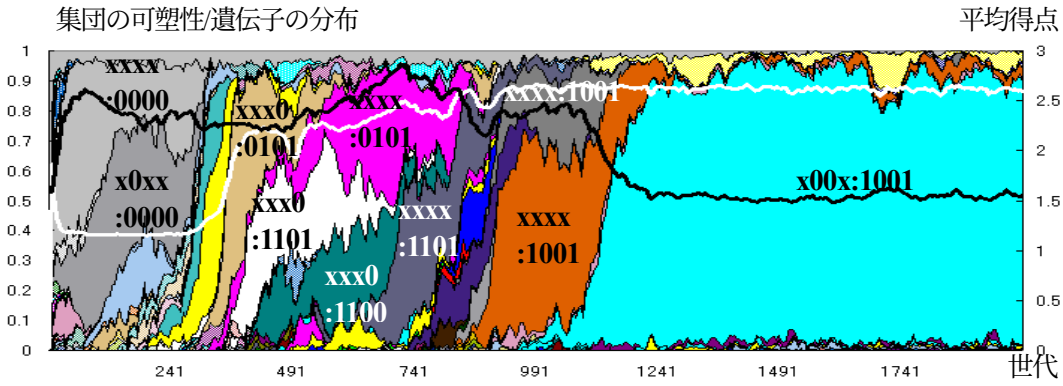

図 7：学習行列を遺伝子に取り込んだ実験結果（[x00x]GT=[1001]に収束した例）

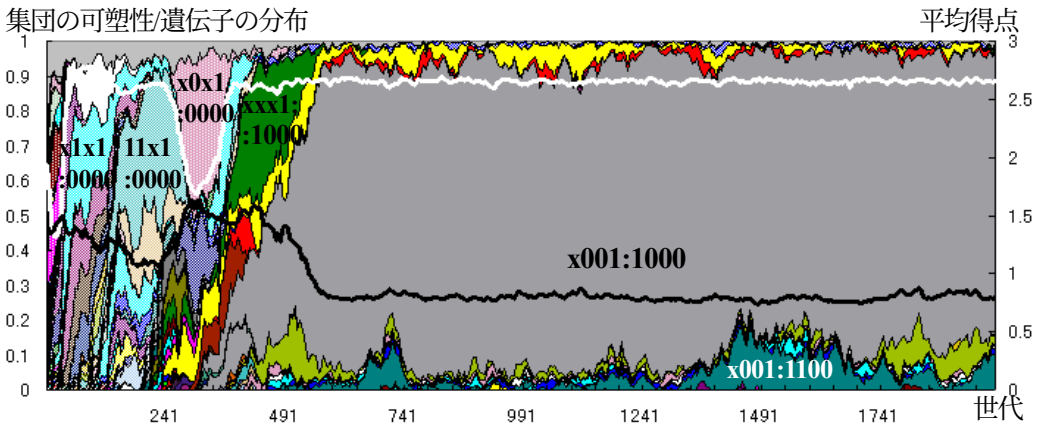

図 8：学習行列を遺伝子に取り込んだ実験結果
（[x001],GT=[1000]に収束した例）

５．まとめ ～工学的応用に向けて～

本研究では，動的環境における進化と学習の相互作用を解析するため，個体間の相互作用のみに依存した動的環境である繰り返し囚人のジレンマゲームの戦略の進化に表現型の可塑性を導入して実験を行った．その結果，このような動的な環境においても Baldwin 効果が観察され，集団は適度な可塑性を持った安定な協調集団へと進化した．

これらの議論はジレンマゲームにおける戦略の進化という，極端に抽象的なモデルにおいての議論であるが，それゆえに多様な展開が可能であると考えている．その一つとして何らかの形で工学的な応用へ展開したいと考えている．

1. 本研究で扱ったメタ・パブロフ学習を最もシンプルな強化学習メカニズムの一つと見なすと，集団全体はマルチエージェント強化学習系として捉えることができる．近年，強化学習に関する研究が盛んに行われている中で，マルチエージェント系における強化学習では，個々のエージェントの学習が互いの環境を不安定にし，学習を困難にさせてしまうことが問題として挙げられる[9]．このような状況において，本研究で採用したような集団全体の学習への依存度は集団の安定性に関する一つの指標となるのではないか．また依存度が進化的に調節されるようなモデリングは有効ではないか．
2. 表現型の可塑性を導入するだけで，学習のメリットやコストを明示的に導入しなくても，進化の過程で双方が創発され，Baldwin 効果を通して遺伝的に取り込むことが可能なメリットは取り込まれ，そうでない部分は学習メカニズムとして残っていくような展開はありえないだろうか．
3. 今回観察された Baldwin 効果は，裏切り集団から協調集団へのシフトの際の 1 回きりであったが，このような進化の過程が繰り返されるうちに，次第に高度な学習メカニズムを支える遺伝的な機構とそれが十分に生かされる集団（環境）が形成されるシナリオがありうるのではないか．このような仕組みをマルチエージェント系における協調作業システムの進化的獲得に展開できないだろうか．

今後はより抽象的なモデルや，コミュニケーションシグナルを通して協調作業を行うようなモデル，トリの鳴き声の性選択学習に関するモデルなどを用いて，進化と学習の相互作用についてさらに解析を進める予定である．

参考文献

[1] Baldwin J. M.: A New Factor in Evolution, *American Naturalist*, Vol. 30, pp. 441-451 (1896).

[2] Turney, P., Whitley, D. and Anderson, R. W.: Evolution, Learning, and Instinct: 100 Years of the Baldwin Effect, *Evolutionary Computation*, Vol. 4, No. 3, pp. 4-8 (1996).

[3] Hinton, G. E. and Nowlan, S. J.: How Learning Can Guide Evolution, *Complex Systems*, Vol. 1, pp. 495-502 (1987).

[4] 鈴木麗璽, 有田隆也: 進化と学習の相互作用－繰り返し囚人のジレンマゲームにおける Baldwin 効果－, 人工知能学会誌, Vol. 15, No. 3.

[5] Axelrod, R.: *The Evolution of Cooperation*, Basic Books, New York (1984).

[6] Lindgren, K.: Evolutionary Phenomena in Simple Dynamics, *Artificial Life II*, pp. 295-311, Addison-Wesley (1991).

[7] Nowak, M. A. and Sigmund, K.: A Strategy of Win-Stay, Lose-Shift that Outperforms Tit-for-Tat in the Prisoner's Dilemma Game, *Nature*, Vol. 364, No. 1, pp. 56-58 (1993).

[8] Boerlijst, M. C., Nowak, M. A. and Sigmund, K.: The Logic of Contrition, *Journal of Theoretical Biology*, Vol. 185, pp. 281-293 (1997).

[9] Tuomas, W. S., Robert, H. C.: Multiagent reinforcement learning in the Iterated Prisoner's Dilemma, *Biosystems* 37, pp. 147-166 (1996).