メタパブロフ：進化と学習による適応を自動調節する繰り返し型囚人のジレンマ戦略

鈴木 麗璽　有田 隆也

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

進化と学習の相互作用に関する重要なトピックに Baldwin 効果がある。本研究では、個体間の相互作用のみを考慮した最適解が決まらない動的環境として、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、戦略にメタパブロフと呼ぶ学習方式に基づく表現型の可塑性を導入し進化実験を行った。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の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、最終的に集団中の大半を占めたメタパブロフ[x00x]型戦略について解析を行い、可塑性を有効に利用した安定な戦略であることを示す。

Meta-Pavlov: Strategies that Self-Adjust Evolution and Learning Dynamically in the Prisoner's Dilemma Game

Reiji SUZUKI　Takaya ARITA

Graduate School of Human Informatics, Nagoya University
reiji@info.human.nagoya-u.ac.jp　ari@info.human.nagoya-u.ac.jp

The Baldwin effect is known as an interaction between evolution and learning. Our concern is to consider the Baldwin effect in dynamic environment, especially when there is no explicit optimal solution through generations and it only depends on interactions among agents. We adopt the iterated Prisoner's Dilemma as a dynamic environment and introduce phenotypic plasticity to strategies by using a meta-learning rule termed "Meta-Pavlov". In this simulation, the Baldwin effect was observed and finally the strategy [x00x] occupied the most of the population. We have analyzed this strategy which has a modest amount of plasticity, and have shown that it satisfies the ESS (Evolutionarily Stable Strategy) condition and establishes robust and cooperative relationships.

１．はじめに

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[2]がある。この現象は、ラマルク的な獲得形質の遺伝の仕組みが無くても、集団における個体の学習が集団全体の進化に方向性を与え、進化のスピードを促進するというものである。Hinton と Nowlan の先駆的な実験[3]によりこの効果が確認されて以来、多くの研究がなされてきたが、ほとんどの場合、世代を通して最適解が固定された静的な環境を前提としており、動的な環境における Baldwin 効果は未解明であった[8]。

しかし、現実世界において学習が有効に働く状況を考えると、静的な環境よりもむしろ動的な環境においてその効力を発揮することから、動的な環境における Baldwin 効果の解明は重要であると言える。

そこで本研究では、Baldwin 効果に対する一般的な解釈である学習のメリットとコストのバランス[7]に注目し、動的な環境において Baldwin 効果が確認されるかどうか、またこのバランスが進化と学習の相互作用にどのように働くかについて知見を得ることを目的とする。

動的な環境を考える場合、大きく 2 つに分けることが出来る。一つは集団が置かれた環境自体が世代を通して変化し、個体の適応度に影響を与える場合である。

Anderson[1]は、世代を通して最適解が変動する動的な環境において学習が進化に与える影響を量的に解析し、動的な環境においては学習にコストがかかっても学習に依存した集団へと進化することを示した。また、佐々木・所[9]は、ニューラルネットを用いて学習する個体が、世代を通して変化する食べ物または毒を表すビット列の入力を正しく識別するようにニューロンの閾値を学習し、適応度に応じて進化するというシミュレーションを行った。その結果、ダーウィン型の進化システムでは学習を前提として次第に動的環境自体に適応していったと報告しており、両研究とも動的環境における学習の重要性を示したものと言える。

もう一つは集団における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場合である。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。

我々は、後者の典型的な例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採用し、我々の提案するメタパブロフ学習に基づく可塑性を戦略の表現型に与えることで、集団全体の進化の過程に学習が与える影響を明らかにすることを目的として研究を行っている[10]。

本稿では、まず、進化実験において Baldwin 効果が有効に働き、協調的で安定な戦略集団へ進化したことを示し、次に進化の過程で出現し最終的に集団中を占めた安定な戦略であるメタパブロフ [x00x]型戦略について解析する。

2. 研究の背景

2.1 Baldwin 効果

Baldwin 効果とは、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[7]、一般には、以下の 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。

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。

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なら、2 つの段階をへて図 1 のようなカーブを描くと考えられる。

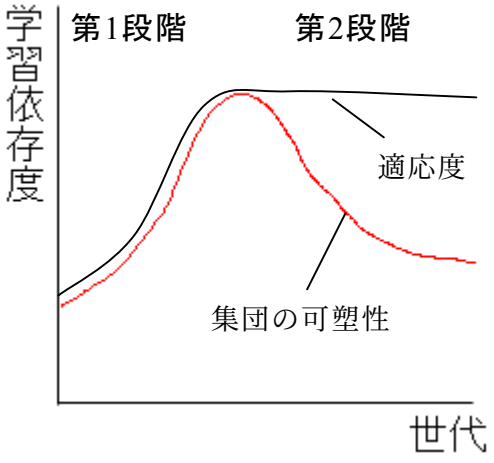


図 1：学習依存度から見た Baldwin 効果

2.2 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、社会的集団に生じるジレンマ状態をシンプルに抽象化したモデルであり、様々な分野で多くの研究がなされているものである。ゲームは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）	(3, 3)	(0, 5)
裏切り（D）	(5, 0)	(1, 1)

（自分の得点, 相手の得点）

もし 1 回きりの対戦なら、裏切りが最良の手であることは自明である。しかし、繰り返し対戦を行う場合には、裏切り合うよりも協調し合った方が双方の利益になるが、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のである。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。

3. 進化実験

以上を踏まえ、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実験を取り上げ、戦略に学習（表現型の可塑性）を導入することで動的環境における進化と学習の相互作用について解析する。

3.1 戦略の遺伝子表現

各個体の持つ戦略を戦略遺伝子列 GS と学習遺伝子列 GL の 2 つの遺伝子列の組で表現する。戦略遺伝子列は Lindgren[4]のモデルと同様な、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。

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

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、

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

と表すことができる。これを戦略遺伝子列とする。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 L_x（0：可塑性を持たない、1：可塑性を持つ）として、学習遺伝子列を

$$GL = [L_0, L_1, \cdots, L_{n-1}] \qquad (3)$$

と定義する。

3.2 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。

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

3.3 学習の例

ここで、メタパブロフ学習の例として、GS=[0001]、GL=[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、学習遺伝子列に"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。

以下、各戦略を、可塑性を持つ学習遺伝子に対応する戦略遺伝子を x と置き換えた戦略遺伝子列でまとめて表現する（e.g. GS＝[1000], GL＝[1001]→[x00x]）。

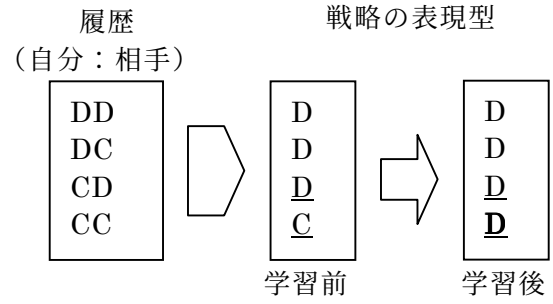


図 2：メタパブロフ学習の例

3.4 繰り返し対戦

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、各戦略個体が意図した手が一定の確率で反転してしまうことである。

繰り返し対戦の一番はじめの手を決定するために、長さ 2 の履歴をランダムに作成し、この履歴に基づいて各戦略個体がはじめの手を決定するものとする。

繰り返しの回数は固定せず、各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、各繰り返し対戦ごとに初期状態（遺伝子表記のままの状態）に戻されるものとする。

3.4 遺伝的オペレーション

上記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。つづいて、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て保存し、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。

4. 実験結果と考察

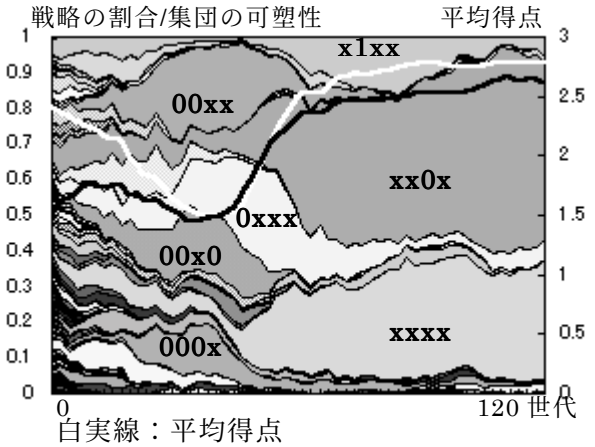
記憶長 2（初期集団はランダム）の集団において、パラメータとして突然変異率 0.001、個体数 1000、ノイズ率 2% 未来係数 0.99、世代数 2000 を用いて進化実験を行った。試行の約 75%で図 3、4 のような集団の可塑性と平均得点の推移の傾向を持つ進化が確認された。

ここで集団の可塑性（黒実線）とは、学習遺伝子列中に占める"1"のビットの割合を示し、これは 2.1 における学習依存度に相当するものと捉えることができる。また平均得点（白実線）は各世代に行われたすべての対戦の得点を平均したもので、互いに協調し合ったときに最も高くなる（3 点）ことから協調の度合いを表すものとして捉える。

この試行における進化の過程の概略を示す。はじめに裏切り的な戦略（[00x0]、[000x]など）が平均得点を低下させた。またそれとほぼ同時に [00xx]、[0xxx]などの可塑的な戦略も集団中を占めた。これらは、同種同士では協調関係を築くことができず、他の裏切り的な戦略に対しても同様に裏切り続けるが、可塑的で協調的な戦略に対してはある程度の協調関係を築く特徴を持ち、その後の可塑的で協調的な戦略の台頭の下地となっていると言える。その後、約 60 世代目から集団の可塑性の増加とともに[xxxx]、[xx0x]、[x1xx]といった可塑的で協調的な戦略が集団中を占め高い平均得

点を持つ集団へと進化した。これまでの過程から、可塑性は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と考えられ、これは Baldwin 効果の第 1 段階と捉えられる。

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、最終的には集団の大部分を[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、可塑性の高い戦略はノイズに対して過敏に反応し、異種との対戦で協調関係を回復するのに時間がかかり、これがコストに相当するため、集団を維持するのに最低限必要な可塑性を持った戦略が選択されたと考えられる。これは Baldwin 効果の第 2 段階と捉えられる。



白実線：平均得点
黒実線：集団の可塑性
色分けされた領域：各世代を占めた戦略の割合
図 3：記憶長 2 での実験結果(120 世代)

5. メタパブロフ[x00x]型戦略の解析

実験結果から最終的に[x00x]型の戦略が集団中のほとんどを占めることがわかった。この[x00x]型戦略はどうして集団中を占めることができたのか、いくつかの角度から解析を行う。

5.1 ESS 条件

集団における戦略が進化の過程で安定であるかどうかの基準として、Maynard Smith が提案した「進化的に安定な戦略（ESS）[5]」がある。集団において ESS を満たす戦略 S の条件は、

E(S, S')を戦略 S と S'との対戦で S が得る得点とすると、

$$E(S, S) > E(S', S) \qquad (4)$$

または

$$E(S, S) = E(S', S) \quad \text{かつ} \quad E(S, S') > E(S', S') \qquad (5)$$

が他のすべての種類の戦略 S' に対して成り立つことである。

[x00x]がこの条件を満たすかどうか確認するために、[x00x]（GS=[0000]、GL=[1001]）と記憶長 2 の可能なすべての戦略 256 個との繰り返し対戦をノイズ率 2%、未来係数 0.99 で 20 回行ったときの各対戦の平均得点を計算した。

図 5 は[x00x]と記憶長 2 の戦略との対戦成績である。横軸は、各戦略の遺伝子列を[GSGL]とならべて 8 ビットの 2 進数として見た場合の値を表す。縦軸は、各戦略個体と[x00x]との対戦において[x00x]の得た

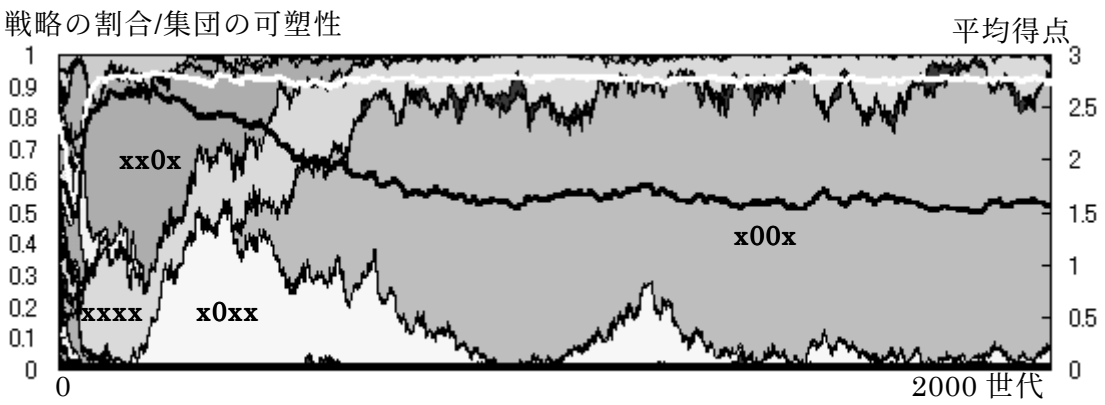


図 4：記憶長 2 での実験結果（2000 世代）

得点と各戦略個体の得た得点の差、すなわち（4）式の左辺と右辺の差である。したがって、[x00x]が ESS であるには縦軸の値がすべて 0 より大であれば良いが、図 4 からこの試行においては ESS 条件を満たしていると言える。

5.2 状態遷移モデル

メタパブロフ[x00x]型戦略は、2 つの可塑性を持った表現型を内部状態とすることで、状態遷移モデルとして捉えることができる。

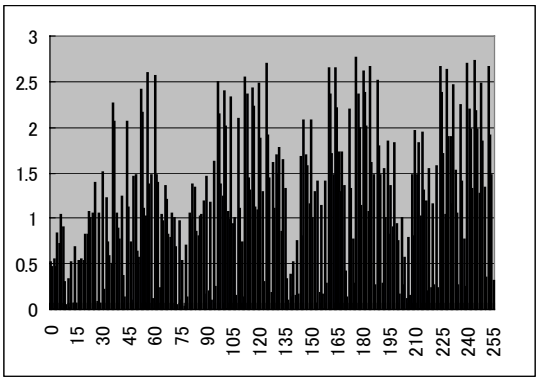


図 5：[x00x]と記憶長 2 の戦略との対戦成績

図 6 に[x00x]型戦略の状態遷移モデルを示す。各矩形は自分と相手の出した手、および可塑性を持った表現型の状態を示す。上下に並んだ“ 0 ”または“ 1 ”は、上が相手の出した手、下が[x00x]の出した手を表す（“ 0 ”=裏切り、“ 1 ”＝協調）。[x00x]の出した手につけられた添え字は、可塑性を持った表現型の現在の状態を示し、順に[x00x]の前の x と後ろの x の表現型の状態を示す。各状態から 2 本ずつ伸びた矢印は、相手の取りうる手に依存して可能な状態遷移先を示している。

任意の戦略と[x00x]との繰り返し対戦は、相手の戦略に応じて矢印を選んでいくことで表現できる。たとえば、[x00x]と全面裏切り戦略[0000]が対戦した場合、相手の手が常に“ 0 ”となるように矢印を選んでいけばよい。この場合、たとえばサイクル②をとり続けることになる。

この図において注目すべき点は、協調関係を持続する状態（状態 A）が崩れたとき、

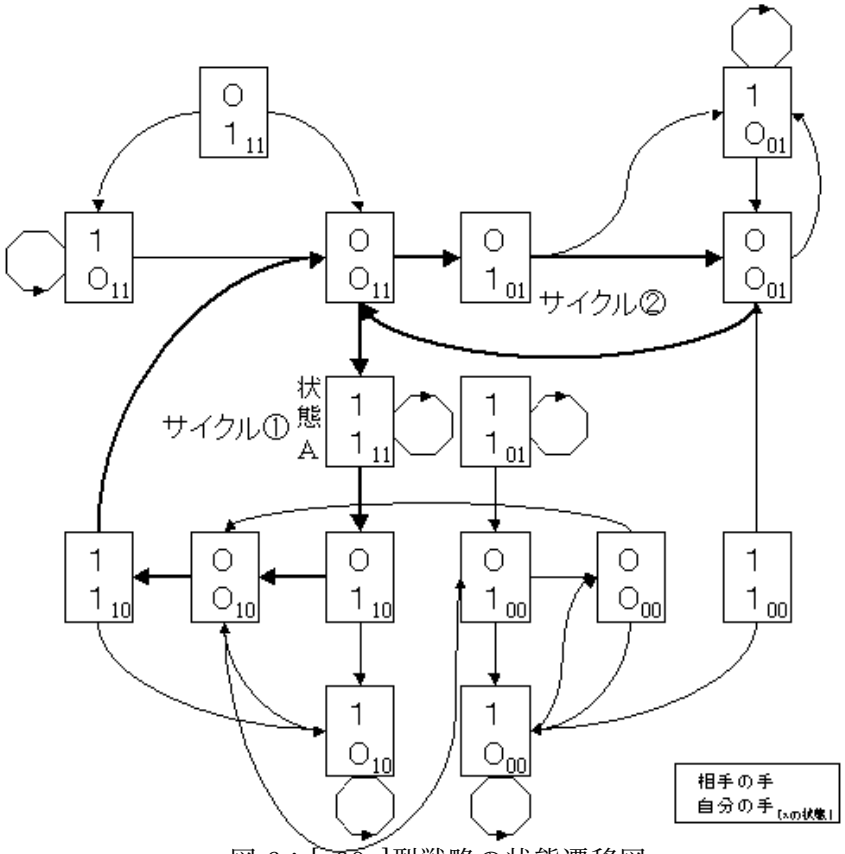


図 6：[x00x]型戦略の状態遷移図

もう一度もとの状態 A に最短で戻るには、相手が「裏切り、協調、裏切り」という複雑な手（サイクル①）を取らなければならないことである。実は、「裏切り、協調、裏切り」のあと協調に戻るという協調の回復過程は、[x00x]同士の対戦において実現される。つまり、協調関係が崩れたときに最も早く協調関係を回復することのできる相手の一つが同種であるということである。逆に言えば、同種以外の戦略に対して、[x00x]は協調関係を回復しにくいことを表しており、この特徴は[x00x]の ESS 的な性質を生み出すメカニズムの一つであるとみなすことができる。

5.3 学習の役割

これまで、進化の過程において、過剰な可塑性が減少した結果、最終的に[x00x]型の戦略が集団中を占めたと主張してきたが、ではどうして、「前の x」と「後ろの x」という 2 つの可塑性を残して減少は止まったのだろうか。状態遷移モデルからわかるとおり、戦略の特徴は遺伝子列全体を総合して決まるが、ここではあえて前の x、後ろの x、その両方それぞれの意義について考察する。

前の x は裏切り中心の戦略に対して、多く点を取らせず、進入させないという点で有効であると考えられる。サイクル②は全面裏切り戦略と対戦した場合にたどるものの一つである。このとき、前の x の可塑性により[x00x]は 2 回に 1 回裏切られる。[x00x]の得る得点は約 0.6 点と低いものの、このおかげで全面裏切り戦略が得る得点は約 2.3 点に下がる（図 7）。[x00x]同士の対戦で [x00x]が得る得点は約 2.8 点であるから、これは[x00x]を進化的に安定な戦略にするのに貢献している。

0000	**000000000000**…平均約 **2.3**
x00x	**010010010010**…平均約 **0.6**

図 7：全面裏切り戦略と[x00x]との対戦例

一方、後ろの x は協調中心の戦略に対して、偶発的な裏切りをきっかけにして裏切りに転じる点で有効であると考えられる。図 6 中央の互いに協調し合った状態から、協調が崩れるとき、後ろの x は 1（協調）から 0（裏切り）に変わる。これは次回協調し合った後、裏切ることを示しており、これは協調を維持するような戦略に対して搾取する（図 8）とともに、相手に点を取らせない点で有効である。

x001	**1111**0**01101101**…
x00x	**1111101001001**…

図 8： [x001]と[x00x]の対戦例
（”0”はノイズ）

以上のとおり、前後の x にはそれぞれ裏切り的・協調的戦略に点を取らせない働きがあると考えられる。しかし、この「他の戦略に対して点を取らせない」働きを生かすには、[x00x]同士の対戦で必ず強固な協調関係が築かれなければならない。前述の通り、[x00x]はサイクル①によってノイズに対してうまく協調を回復するが、この遷移の中では前後の x についてそれぞれ 2 回ずつ交互に学習（結果的に状態が変化しない学習も含む）が行われている（図 9）。これは前後の x による学習が共同してうまく働いた結果、強力な協調関係を築いていることを示している。

x00x	**1111**0**0101111**…
x00x	**111110101111**…

図 9：[x00x]同士の対戦での協調の回復例
（”0”はノイズ）

6. おわりに

本稿では、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、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。このとき、進化の過程で協調的で安定した集団へと直接進化したのではなく、一旦集団の可塑性が増加する方向へ進化し、十分な協調関係を実現してから、可塑性の減少とともに安定な集団へと進化したことが確認された。また、このような個体間の相互作用に依存した環境においては、協調関係といった集団レベル

の形質の獲得について Balwin 効果が働くことが確認された。

さらに、最終的に集団の大部分を占めたメタパブロフ型[x00x]戦略の解析を行ったところ、[x00x]型戦略は必要最低限の可塑性を持った強力な戦略であることが判明した。これは[x00x]型の戦略が純粋にジレンマゲームにおける強力な戦略として興味深いだけでなく、このような進化と学習の割合が自動的に調節される進化の枠組みが、動的な環境に対する集団の安定性を実現するための重要なファクターとなっていることを示唆する点で意義深いと考えられる。現在、[x00x]型戦略について解析を続けるとともに、記憶長や学習方式を固定しないオープンエンドな進化実験などを行っている。また、このような進化と学習の割合が相互作用する仕組みの工学的応用についても検討している。

参考文献

[1] Anderson R.W. (1995): Learning and Evolution: A Quantitative Genetics Approach, Journal of Theoretical Biology, 175, pp. 89-101.

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

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

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

[5] Maynard Smith. J. (1982): Evolution and the Theory of Games, Cambridge University Press.

[6] Nowak. K and Sigmund K. (1993): A Strategy of Win-Stay, Lose-Shift that Outperforms Tit-for-Tat in the Prisoner's Dilemma Game, Nature, 364, pp. 56-58.

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

[8] 有田隆也 (1999): 「人工生命」, 科学技術出版.

[9] 佐々木貴宏, 所 真理雄(1997):進化的エージェント集団の動的環境への適応, コンピュータソフトウエア, Vol. 14, pp. 33-46.

[10] 鈴木麗璽, 有田隆也(1999): 囚人のジレンマゲームにおける Baldwin 効果,人工知能学会第 13 回全国大会論文集.