Evolutionary Differentiation of Learning Abilities – a case study on optimizing parameter values in Q-learning by a genetic algorithm

Tatsuo Unemi[*], **Masahiro Nagayoshi**, **Nobumasa Hirayama**,
Toshiaki Nade, **Kiyoshi Yano**, and **Yasuhiro Masujima**
Department of Information Systems Science, Soka University
1-236, Tangi-machi, Hachioji, 192 Tokyo, JAPAN
e-mail: {unemi,rorry,nobu,toshi,papino,yas}@iss.soka.ac.jp

Abstract

This paper describes the first stage of our study on evolution of learning abilities. We use a simple maze exploration problem designed by R. Sutton as the task of each individual, and encode the inherent learning parameters on the genome. The learning architecture we use is a one step Q-learning using look-up table, where the inherent parameters are initial Q-values, learning rate, discount rate of rewards, and exploration rate. Under the fitness measure proportioning to the number of times it achieves at the goal in the later half of life, learners evolve through a genetic algorithm. The results of computer simulation indicated that learning ability emerge when the environment changes every generation, and that the inherent map for the optimal path can be acquired when the environment doesn't change. These results suggest that emergence of learning ability needs environmental change faster than alternate generation.

1 Introduction

There are many layers in the Artificial Life researches such as molecular dynamics, evolution, development, learning, collective behavior, and so on. One of the methods for fruitful studies is on combination of two or three of these layers, such as evolutionary development system, collective behavior of learners. We already finished the first stage of above two kinds of combinations (Nade et al, 1994, Unemi, 1993). On the third combination, evolution of learners, Todd and Miller have been pursuing evolutionary process to organize associative neural networks that learn by a simple Hebbian rule (Todd and Miller, 1990). Ackley and Littman mentioned genetic acquisition of evaluation network in a neural network based reinforcement learning method (Ackley and Littman, 1992) and a distributed Lamarckian evolution (Ackley and Littman, 1992). Their studies provided us fruitful suggestion to understand a process of emergence of intelligent creatures. However, we are just starting our steps toward the real intelligent creature, that is, human. We need more studies to complement the pioneers' works and to advance our understanding in this field.

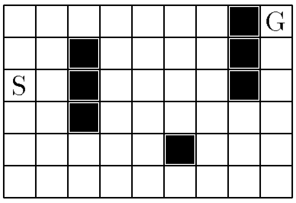


Figure 1: An example of the maze. Black squares are obstacles. The letter S and G indicate the start position and the goal position respectively.

This paper describes the first stage of our study on evolution of learning abilities of which final goal is to make learning ability to emerge structurally. As the first step of this challenge, we tried to optimize the inherent learning parameters of a reinforcement learning mechanism using a genetic algorithm. We can intuitively expect that it is unnecessary for creatures to have learning abilities if the environment is stable and doesn't change through many generations but learning abilities are needed if the environment changes every generation. To realize these phenomena on the computer, we designed a simulator of evolvable learners that consists of a simple learning task, a simple learning algorithm, and a simple genetic algorithm.

The rest sections of this paper describe the specifications of the simulator, the experiments, and their results.

2 Task for individual

The task for each individual is a simple maze exploration designed by Sutton as a testbed for his Dyna learning architecture (Sutton, 1990). The maze an individual creature lives in is a two-dimensional grid world of nine columns and six rows. Some of the grids are occupied by obstacles so the individual cannot position there. Figure 1 shows a typical map of the maze used in the experiments described later.

In each execution step, it moves from the current position to the adjacent grid in the maze not occupied by an obstacle. From fixed start position, it explores the maze

[*] Soka University, and Laboratory for International Fuzzy Engineering Research, Siber Hegner Bldg., 3rd Fl., 89-1 Yamashita-cho, Naka-ku, 231 Yokohama, JAPAN

toward the fixed goal position. It gets positive reinforcement signal when it reaches the goal and then starts exploration from the start position again. It knows its position in each step.

This is a typical task for simple reinforcement learning, where the environment is stable, deterministic, and discrete. Some readers may feel that such settings are too simple as a model of life, but it would be better to start a simple model as possible at the first stage.

3 Learning algorithm

We employ a one step Q-learning using look-up table proposed by Sutton (1990), because its algorithm is very simple and the look-up table has an ability to represent a map of the world. A brief description of this learning algorithm is as follows.

In each execution step, the learner gets sensory input from the environment and decides its action based on the input data. Then it takes the action, and receives reinforcement signal if available, and then goes to the new state. The look-up table contains the Q-values that estimate expected reward corresponding to the all of possible states and actions, that is, Q-value Q_{xa} indicates expected reward when the individual takes the action a at state x. For the task we use, the look-up table contains $4 \times 2 + (7 + 4) \times 2 \times 3 + 4 \times 7 \times 4 = 186$ Q-values if it includes no obstacle, because of two actions in four corners, three actions in four sides, and four actions in inner grids.

The action is determined in each step according to a probability of Boltzmann distribution. The probability for selecting action a at state x is defined as

$$P(a|x) = \frac{\exp(\alpha Q_{xa})}{\sum_{j \in Possible\ Actions} \exp(\alpha Q_{xj})}$$

where α, exploration rate, is the inverse value of temperature. When the value of α is large, it tends to select the action of maximum Q-value rigidly. When the value of α is small, nearly zero, it tends to take a random action regardless of the Q-values.

When it selected action a at state x, the result state was y, and it received reward r, then the Q-value Q_{xa} is revised by the assignment equation:

$$Q_{xa} \leftarrow Q_{xa} + \beta\,(r + \gamma \max_b Q_{yb} - Q_{xa})$$

where β is learning rate and γ is discount rate of rewards.

The learner explores the maze according to its action selection mechanism and estimates the value of each action at each state only referring to delayed rewards through its own experience. Each individual creature lives alone, that is, the individuals in population have no interaction with each other.

Learning performance strongly depends on the values of parameters, especially on the initial Q-values, though theoretical guarantee of convergence to the global optimum solution after infinite times of trials is proved by Watkins (1993). Because the value of β should gradually increase for certain convergence, we set up the start value and the end value of α and β, and gradually change the values step by step under the constant difference.

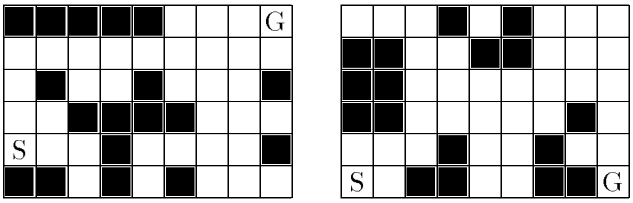


Figure 2: Examples of the random mazes.

According to our preliminary experiments in the above problem, when all of the initial Q-values is set to zero, 4,000 steps provides an enough number of trials to find an optimal path, sometimes global optimum but often local optimum. So, the life span of each individual is set to 4,000 steps fixed in the experiments described later.

4 Genetic algorithm

This section describes the genetic algorithm we use here.

4.1 Genetic code

All of inherent learning parameters have continuous numerical values. So, we employ eight bits integer coding for each parameter and translate each genetic code to real number of its range proportionally, for example, because the range of γ is $[0, 1]$, the value of γ is set to $G_\gamma/255$ where G_γ is the value of eight bits unsigned integer on the gene corresponding to γ. The ranges of Q_{xa}, β, and γ are $[0, 1]$, and the range of α is $[0, 63.75]$ $(63.75 = 255/4)$. One genome contains 221 bytes, $6 \times 9 \times 4 = 216$ Q-values, α_{start}, α_{end}, β_{start}, β_{end}, and γ. The reason we use redundant 216 Q-values rather than 186 values as described above is merely to simplify the implementation on the computer program.

4.2 Fitness

The fitness of each individual is calculated by counting the times the learner achieves at the goal position in its later half of life. The life span is set to 4,000 steps as described above, counting is done during the later 2,000 steps. Because the first half of life should be treated as a moratorium, we ignore the performance of that period.

4.3 Genetic operations

We use a simple genetic algorithm with a ranking and elitist strategy using selection, crossover, and mutation. The following list summarizes the algorithm.

Selection Remain the best third of genomes to the next generation.

Crossover Replace the middle third of genomes to the genomes made by crossover operation between each one of the best third and the middle third. Each genome includes two pieces of byte string chromosomes where the one contains Q-values and the other one contains the five parameters. One point crossover is applied to each chromosome independently not bitwise but bytewise.

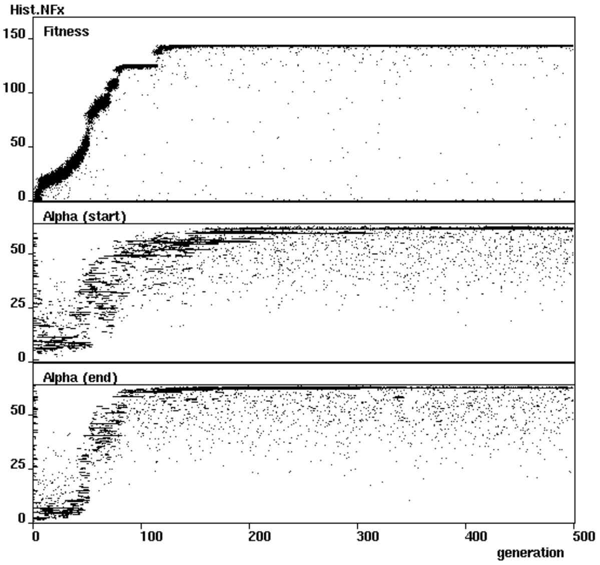


Figure 3: An evolutionary process without learning on fixed environment.

Mutation Replace the worst third of genomes to the mutant of the best third. One byte randomly selected from each chromosome is modified by pseudo-Gaussian distribution.

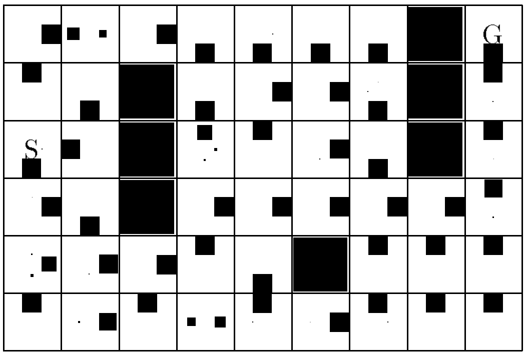
Starting from the population of randomly initialized genomes, evaluation and genetic operation cycles are iterated. The result of learning by an individual doesn't inherit to the offspring but only the natural characteristics, that is, not Lamarckian but Darwinian evolution.

5 Experiments

For the purpose of this research, we tried experiments concerning with four kinds of settings as follows.

1. Without learning but only evolution on the fixed environment.
2. Without learning but only evolution on the environment that changes randomly every generation.
3. With learning and evolution on the fixed environment.
4. With learning and evolution on the environment that changes randomly every generation.

"Without learning" means the individual doesn't learn in any trial, that is, Q-values are not modified. The behavior of the agent without learning depends only on its inherent Q-values and α. The values of β and γ have no effect. The map of fixed maze is as shown in Figure 1 above. A random maze is designed as the start position is located in the left most column and the goal position is located in the right most column and a path from start to goal exists. Figure 2 shows two examples of the random mazes.



$\alpha_{\text{start}} = 61.0, \alpha_{\text{end}} = 62.0,$
$\beta_{\text{start}} = 0.24, \beta_{\text{end}} = 0.86, \gamma = 0.85$

Figure 4: Probabilities of action selection of the best individual at 500th generation – the case without learning on fixed environment. The size of black square in each grid represents the probability of selecting the action at that state, that is, $P(a|x)$.

We expected that it needs learning abilities if the environment changes every generation but it doesn't need if fixed. Setting 1 was expected to lead the acquisition of inherent map of the environment, and setting 4 was expected to lead the acquisition appropriate parameter values for learning. In setting 2, it seemed to be hard to adapt the environment. However, we could not predict what would happen in setting 3.

Using the following values of genetic parameters, we examined ten runs for each setting with distinct random number sequences.

Population size	=	100	individuals
Life span	=	4,000	steps
Number of generations	=	500	generations

The rest of this section describes a summary of the experimental results.

5.1 Without learning on fixed environment

Figure 3 shows a typical trace of an evolutionary process. As shown in Figure 4, the optimum path of the maze is acquired through a punctuated equilibrium evolution, where Q-values represent the map of the path and the value of α becomes not always the maximum but large enough to make selection rigid. All of ten runs shows the similar pattern of evolutionary process, but some of them were trapped at a local optimum even at 500th generation.

5.2 Without learning on changing environment

As shown in Figure 5, it is difficult to adapt the environment to walk through the optimal path. However, from Figure 6, we can observe that the agent acquired the tendency to go right because the start position is located at the left most column and the goal position is located at the right most column. All of ten runs shows the similar pattern of evolutionary process. This fact suggests that it is difficult to adapt to any changing environment without learning ability.

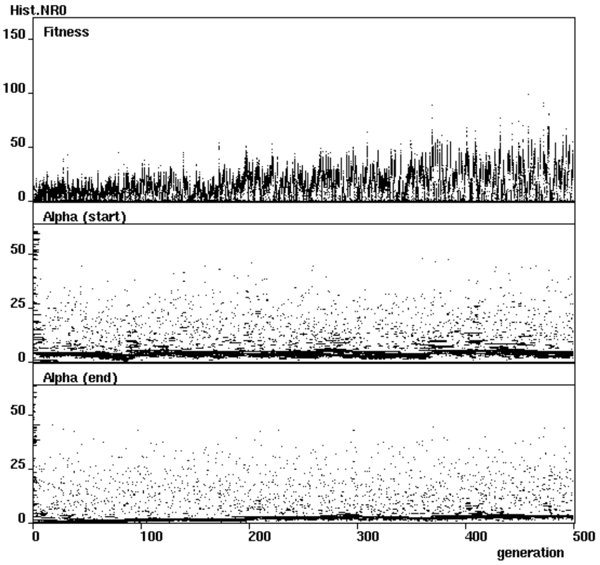


Figure 5: An evolutionary process without learning on changing environment.

5.3 With learning on fixed environment

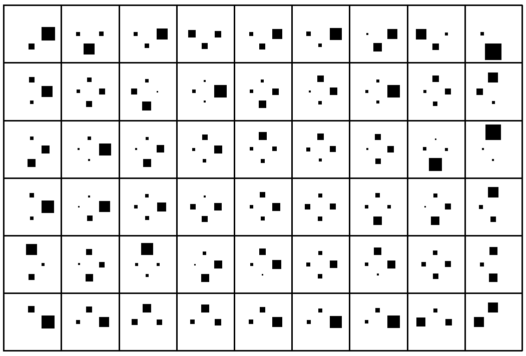
We observed both cases where the agents with efficient learning abilities appeared and where the agents who inherently know the optimal path appeared. Figure 7 and 8 show the former case, and Figure 9 and 10 show the later case. The typical difference between these two cases is seen at the value of α_{start}. In the former case, α_{start} converged low that leads the agent to radical exploration at young age. This characteristic is useful for learning. In the later case, both α_{start} and α_{end} are high that makes the agent to take a conservative action following the inherent Q-values. The high value of α prevents the agent from learning. This means that the value of β and γ make no effect in this case.

5.4 With learning on changing environment

As we expected, the agent with an efficient learning ability emerged. Figure 11 shows a typical pattern of the evolution, where fitness values widely vary because the length of the optimal path of each maze randomly generated is also widely varies. The value of α_{start} becomes low and the value of α_{end} becomes high as similarly as the case of fixed environment where learning abilities are acquired.

6 Conclusion

We investigated the relation between learning and evolution under the different stability of the environment. As we expected before experiments, learning abilities emerged when the environment changed every generation, and inherent knowledge about the world was acquired when the environment fixed. However, we are surprised at the observation that learning ability could emerge even when the environment fixed. We guess that this phenomenon happens because of our assumption of genetic codes that only include some numerical parameters but do not mention any structural information of learning mechanism. The observation that to change the environment every generation causes evolutionary acquisition of efficient learning abilities suggests us that the condition of environmental change strongly effects the structural emergence of learning abilities.



$\alpha_{\text{start}} = 4.8, \alpha_{\text{end}} = 2.2,$
$\beta_{\text{start}} = 0.64, \beta_{\text{end}} = 0.37, \gamma = 0.82$

Figure 6: Probabilities of action selection of the best individual at 500th generation – the case without learning on changing environment.

The results complement the work on genetic acquisition of evaluation network by Ackley and Littman, because we here focused on differentiation of adaptation strategies, evolution versus learning. The range of learning rate β doesn't include negative value, so the learner always mentions the reinforcement signal at the goal position as reward, never as punishment. It is one of quick experiments to make the range of β to include negative value so as to consider acquisition of evaluation function. Both of the experiments by Todd and Miller, and Ackley and Littman, did not mention the relation between the stability of the environment and evolutionary process of learning abilities. The experimental results can provide an extension of our knowledge in this field combining with the pioneers' results.

Starting from the first stage of the research described here, our future work will include more precise investigation of the effects of genetic parameters and development of a methodology of structural emergence of learning abilities.

Acknowledgment

The authors would like to thank Shigenobu Kobayashi, Masayuki Yamamura, Norihiko Ono, Hiroshi Deguchi, Hisashi Tamaki, Osamu Sakura, and Kazuhiro Matsuo for their useful suggestions on earlier work of this research. The authors would also like to thank the referees for their helpful comments for the manuscript of this paper, and thank Toshiro Terano at Laboratory for International Fuzzy Engineering Research for the support of this research.

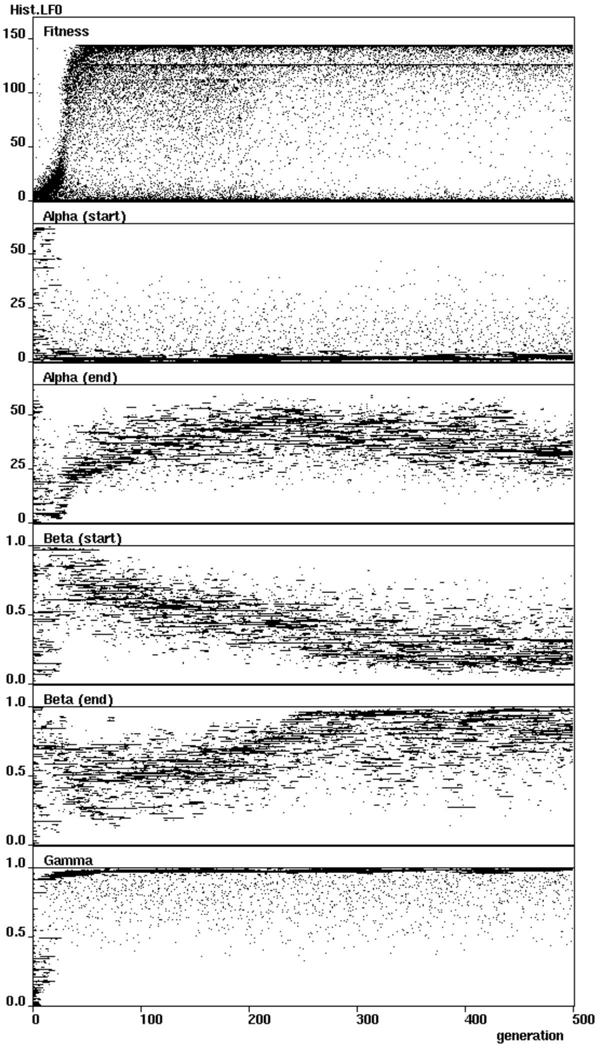
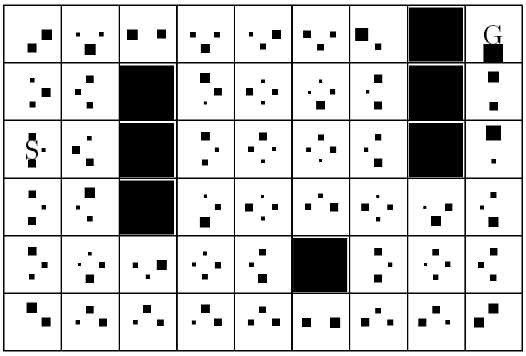


Figure 7: An evolutionary process with learning on fixed environment – 1.



$\alpha_{\text{start}} = 2.8, \alpha_{\text{end}} = 36.5,$
$\beta_{\text{start}} = 0.25, \beta_{\text{end}} = 0.81, \gamma = 0.98$

Figure 8: Probabilities of action selection of the best individual at 500th generation – the case with learning on fixed environment – 1.

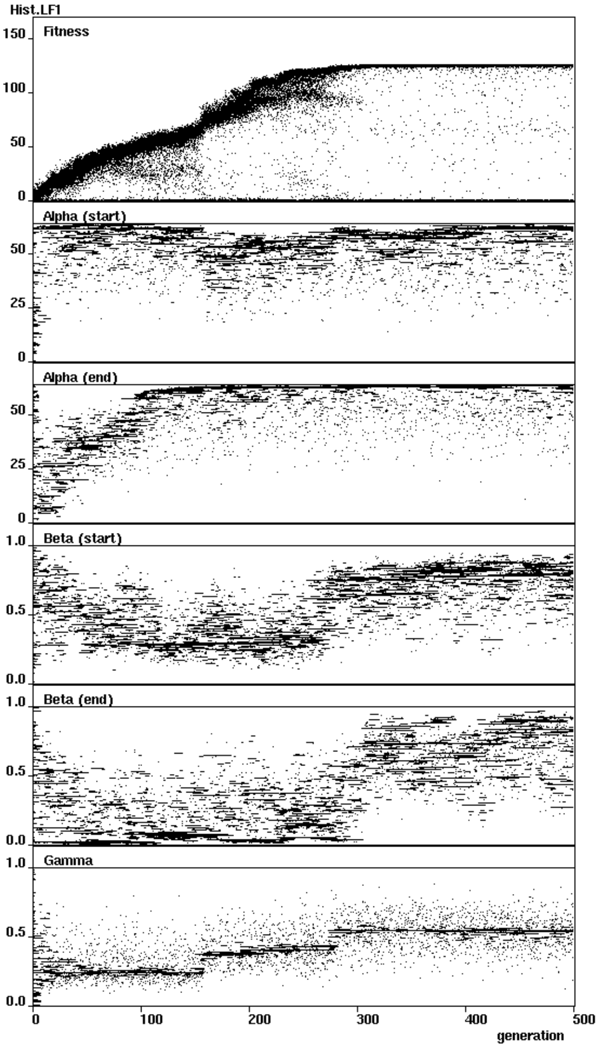
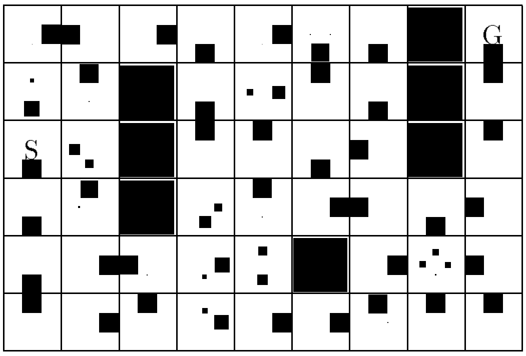


Figure 9: An evolutionary process with learning on fixed environment – 2.



$\alpha_{\text{start}} = 61.2, \alpha_{\text{end}} = 63.0,$
$\beta_{\text{start}} = 0.80, \beta_{\text{end}} = 0.92, \gamma = 0.55$

Figure 10: Probabilities of action selection of the best individual at 500th generation – the case with learning on fixed environment – 2.

References

Ackley, D. and M. Littman. 1992. Interactions between Learning and Evolution. In *Artificial Life II*, edited by C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen. Addison-Wesley, 487–509.

Ackley, D. and M. Littman. 1992. A Case of Distributed Lamarckian Evolution. Presented at *the Third International Workshop on Artificial Life*, Santa Fe, NM.

Nade, T., M, Nagayoshi, N, Hirayama, Y. Masujima, K. Yano, and T. Unemi. 1994. A Simple Development System on 3D Euclidean Space and its Evolution. *IPSJ SIG Notes* 94:20:25–30, in Japanese.

Sutton, R. S. 1990. Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming. *Proceedings of the Seventh International Conference on Machine Learning*, 216–224.

Todd, P. M. and G. F. Miller. 1990. Exploring Adaptive Agency II: Simulating the Evolution of Associative Learning. *From Animals to Animats – Proceedings of the First International Conference on Simulation of Adaptive Behavior*, 306–315.

Unemi, T. 1993. Collective Behavior of Reinforcement Learning Agents. *Proceedings of the 1993 IEEE/Nagoya University WWW on Learning and Adaptive System*, 92–97.

Watkins, C. J. C. H. and P. Dayan. 1993. Technical Note Q-Learning. In *Reinforcement Learning*, edited by R. S. Sutton, Kluwer Academic Pub., 55–68.

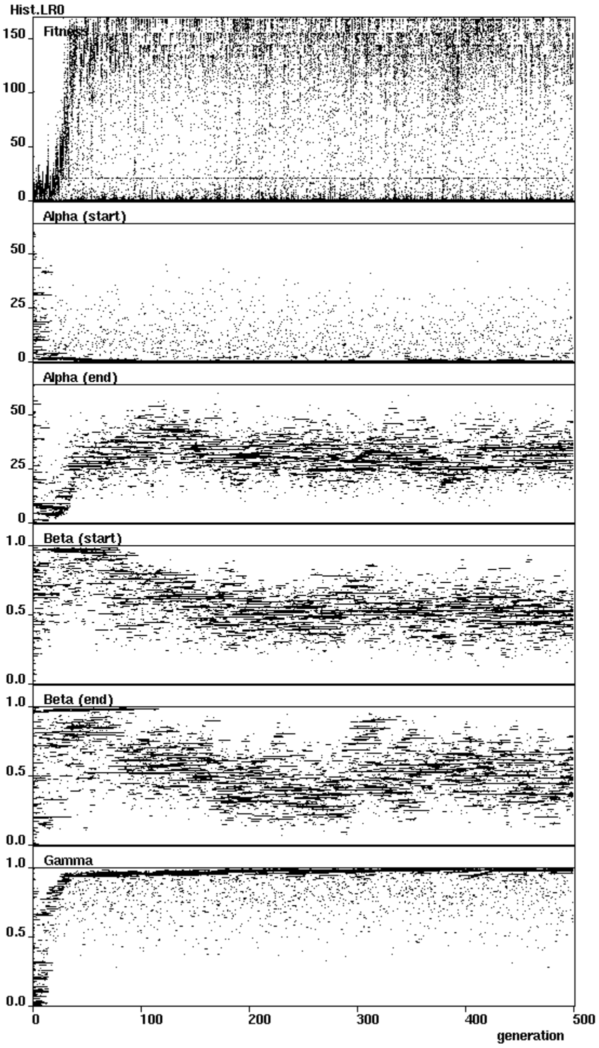


Figure 11: An evolutionary process with learning on changing environment.