Introns for Accelerating Quasi-Macroeolution*

Ken NAITOH**

Computational experiments are performed with the simple genetic algorithm (SGA) for a wide range of intron lengths, crossover rates, mutation rates, and replication rates. The best-adapted species emerges widely with a critical intron length proportional to the overall crossover rate in an environment of discontinuous evolution (quasi-macroeolution). This result coincides with the relation between the intron length and mating rate observed in actual Eukaryotes. On the basis of this knowledge, a method for optimizing the crossover rate in the SGA is proposed for the purpose of accomplishing artificial macroevolution in engineering problems.

Key Words: Optimum Design, Probabilistic Method, Artificial Intelligence, Introns, Quasi-Macroeolution, Simple Genetic Algorithm

1. Introduction

The applications of genetic algorithms (GAs)⁽¹⁾ have been seen as a promising approach to the optimization of various types of network systems, such as electronic circuits and mail delivery routes, as well as the shapes of machines like airfoils and engines. Through the use of an artificial brain⁽³⁾, including a memory system such as an artificial neural network and a learning theory like the genetic algorithm, researchers have tried to construct autonomic control systems such as those for automobiles. However, the parameters used in GAs have so far been chosen on a trial-and-error basis, since theoretical considerations of GAs are still at an immature stage.

One of the starting points for understanding the simple genetic algorithm (SGA)⁽¹⁾ is the idea of degeneracy and redundancy^{(3),(4)}. The quasi-species model based on this idea⁽⁴⁾ provides a conceptual explanation for the diversity of species in evolutionary strategy⁽⁵⁾ but cannot explain the artificial evolution in the SGA because the quasi-species model is limited to cases without a crossover operator. Vose⁽⁶⁾ proposed a

theoretical model for the SGA with a quasi-linear replication operator. However, since the model is constructed as a dynamical system in an extremely high-dimensional hyperspace, it is difficult to obtain a solution to the model with today's computing power. It is also difficult to use the model to answer such essential questions as what kind of characteristics underlying the environment influence the diversity of species.

In the previous reports^{(7),(8)}, a space for clarifying the environment was proposed in the form of a degeneracy structure (DS)⁽⁸⁾. A macroscopic kinetic equation system, using only four variables to describe the dynamical behaviors underlying the SGA, was also proposed in the space of the DS. With an extremely short calculation time in comparison with that of the SGA, the proposed kinetic equation can predict the time-dependent features of the SGA quantitatively for a wide range of parameters such as crossover rates, mutation rates, replication rates, population size, string lengths, initial conditions, and problem differences. The proposed kinetic equation may be used as an artificial intuition model in order to infer how to optimize the parameters in the SGA. However, the kinetic equation is not useful in cases where introns^{(9),(10)} exist in genes⁽⁸⁾.

This report describes a large number of

* Received 13th June, 1997

** Nissan Research Center, 1 Natsushima-cho, Yokosuka-shi, Kanagawa 237-8523, Japan

computational experiments performed with the SGA for a wide range of intron lengths, crossover rates, mutation rates, replication rates, and environments and the function of introns is clarified. A method is then proposed for optimizing the crossover and replication rates in the SGA with introns. Emphasis is placed on the fact that the probability at which artificial macroevolution can be realized increases by using long introns or by reducing the overall crossover rates for the problems of discontinuous evolution (quasi-macroevolution).

These computational experiments also shed important light on the function of introns in actual Eukaryotes.

2. A Computer Model of the Intron-Exon System

The SGA employed here is based on the following assumptions.

(1) Gene

A gene is defined as a one-dimensional string having a constant string length L equal to the number of loci in the gene, with each locus l of the gene I having the value $e_I(l)$ of 1.0 or 0.0. The string length L is independent of time t . The number R of all the possible types of strings is 2^L .

(2) Population Size

Let N be the population size of genes in each generation (time). The population size is assumed to be constant for simplicity.

(3) Fitness Value and Best-Adapted Species

Fitness value f_I of the gene I determines the replication speed, the probability at which offspring are replicated during one generation. The best-adapted species is defined as the gene that has the highest fitness value among all the possible types of strings.

(4) Exons and Introns

Introns in genes are regions that are not expressed, or intragenic regions, while regions of chromosomes that are expressed as proteins are called exons^{(9),(10)}. Thus, in the SGA, each intron locus is independent of the fitness value, although each exon locus can vary the fitness value of a gene. The numbers of exon and intron loci are denoted as Le and Li . $Le + Li$ is equal to L . In the present report, exons exist from locus 1 to locus Le , while introns exist from locus $Le + 1$ to locus L .

(5) Crossover Operation

A one-point crossover operator⁽¹⁾ is used. The members of the strings are first mated at random. Each gene has the same probability of being selected for crossover. The number of the mated pairs divided by the population size N is the overall crossover rate α_c . The rate α_c is independent of time. Each pair of strings undergoes a crossover as follows: an integer

position l_{cross} along the string is selected at random between 1 and the string length L ; two new strings are created by swapping all characters between 1 and l_{cross} inclusively.

(6) Mutation Operation

Single-bit mutation⁽¹⁾ is used. Each gene has the same probability of undergoing mutation. An integer position l_{mut} along the string length is first selected at random between 1 and the string length L . Then, a value $e_I(l_{mut})$ at the locus l_{mut} is replaced with $1 - e_I(l_{mut})$. The number of the mutated strings divided by the population size N is the overall mutation rate α_m . The rate α_m is independent of time.

(7) Replication Operation

An on-off type of replication operator is used to replicate offspring according to the fitness value f_I . First, we consider the expected value of the fitness value, $Y(t) = \sum_{I=1}^N f_I / N$. A threshold value K , independent of time, is then introduced as an artificial constant for controlling the overall replication rate. Then, the information of each gene having a fitness value larger than the critical fitness value $Y(t)N/(2K)$ is copied to another gene selected randomly. The present threshold value gives the edge of death.

Note that the overall replication amount in each generation generally maintains a constant level, since the critical fitness value $Y(t)N/(2K)$ becomes larger when the expected value increases. It should be noted that a large threshold value implies a high replication rate system, since the critical fitness value $Y(t)N/(2K)$ becomes small for high threshold values.

(8) Initial Gene Pool

The initial genes of the first generation are produced by using a random number generator.

(9) Concrete Problem (Environment)

In strict terms, the environment, the difference between problems, can be defined by the fitness value distribution of all the species in the sequence space⁽⁴⁾. However, because the sequence space is a hyperspace, it is difficult to clarify the environment. Therefore, in order to classify environments macroscopically, the underlying degeneracy structure (UDS) was proposed, which is the appearance probability distribution of all the possible genes plotted on the two-dimensional surface spanned by the Hamming distance (HD) and the fitness value^{(7),(8)}. In cases where introns do not exist, the UDS clearly expresses the differences between problems. In cases where introns exist, the differences between problems can be clarified by the UDS for the gene system from which introns have been eliminated⁽⁸⁾. (It was also proposed that the initial degeneracy structure (IDS), the appearance probability distribution of the genes in the

first generation plotted on the two-dimensional surface spanned by the Hamming distance (HD) and the fitness value, can be used to infer the environment, instead of the UDS⁽⁸⁾. This is because we can not usually see the UDS before performing the optimizations^{(7),(8)}.

In the previous reports^{(7),(8)}, problems were classified into two categories, namely microevolution and quasi-macroevolution, on the basis of the UDS. When all genes belonging at least to one class of the Hamming distance in exons have a value of $f=0$, the UDS is defined as an environment of quasi-macroevolution. (There are Le classes of the Hamming distance in the gene system of the exons length Le .) When there are genes of $f>0$ in each class, the UDS is defined as an environment of microevolution. Then the problems with the UDS of microevolution were thoroughly analyzed in the previous work^{(7),(8)}. Thus, in the present study, the problem of quasi-macroevolution, a deceptive problem⁽¹⁾, will be analyzed.

The fitness value f_i is evaluated by the equations

$$f_i = \bar{f}_i, \text{ if } HD_i = 0 \text{ or } HD_i \geq L_c + 1, \quad (1.a)$$

$$f_i = 0, \text{ if } 1 \leq HD_i \leq L_c, \quad (1.b)$$

$$\bar{f}_i = \sum_{l=1}^{Le} V(l) e_i(l), \text{ if } \sum_{l=1}^{Le} W(l) e_i(l) \leq TW, \quad (2.a)$$

and

$$\bar{f}_i = 0, \text{ if } \sum_{l=1}^{Le} W(l) e_i(l) > TW, \quad (2.b)$$

where HD_i denotes the Hamming distance of exons between the gene i and the best-adapted species⁽⁴⁾, the weight W and the value V take randomly generated integer values between 1 and 30, and the limitation TW is set to be 140 in this study. (The limitation TW implies the total weight limitation in the knapsack problem⁽¹¹⁾.) The concrete values of L_c are given in the following sections. The best-adapted species is only one among all the possible types of strings when introns are not used and the number of the best-adapted species is 2^{Li} when introns exist. The best-adapted species in this problem was calculated previously by the all-searching procedure.

Here, the genes with structures close to the best-adapted species are valueless because of Eq. (1.b), as is shown conceptually in Fig. 1. Therefore, this is the environment of quasi-macroevolution. (The system of Eqs. (1.a), (1.b), (2.a) and (2.b) is a deceptive problem⁽¹⁾ in informatics, although Eqs. (2.a) and (2.b) only are a knapsack problem⁽¹¹⁾.)

This deceptive problem of Eqs. (1.a), (1.b), (2.a), and (2.b) is an artificial deceptive problem. However, some actual problems, such as ones involving network pattern optimization in electronic circuits have the environment of this deceptive problem.

Fig. 1 Environment for quasi-macroevolution

The evolutionary route is varied by adjusting six artificial inner parameters, α_c , α_m , K , N , Le and Li , and is also influenced by an external parameter, the initial appearance probability distribution of genes.

3. Computational Experiments for Various Intron Lengths

The system of $Le=16$ and $Li=0$ was examined thoroughly in the previous reports^{(7),(8)}. In the present study as well, the exon length Le is fixed to be 16 and the intron length Li is varied from 0 to 484. In this section, the overall crossover rate α_c is first set at 0.50, which is a medium value, and mutation is not used. The threshold value K is set to be 400, which is optimized for the representative environments^{(7),(8)}. The population size N is set to be 1 000, which is fairly small in comparison with the number of all the possible types of strings. In this section, the environment is fixed by setting the critical Hamming distance L_c in Eqs. (1.a) and (1.b) to be 3.0.

Three computer simulation trials are performed with the SGA for each intron length of $0 < Li < 485$. Then, fifty trials are performed for $Li=0$. The total number of trials is 1 502.

Figure 2 shows the appearance probability of the best-adapted species in the 500th generation as a function of the intron length Li . It is seen in Fig. 2 that, when the intron length exceeds a critical value, the best-adapted species emerges widely before the 500th generation.

Figure 3 shows the representative time-dependent appearance probabilities of the best-adapted species (Group A), the mutants of $0 < HD \leq 3$ (Group B), and the other mutants of $HD \geq 4$ (Group C) for various intron lengths. The SGAs of $Li < 24$ can never find the best-adapted species, although the SGAs of $Li \geq 24$ can produce a rapid population increase in the

Fig. 2 Appearance probability of the best-adapted species at the 500th generation as a function of the intron length, in the environment for quasi-macroevolution of $Lc=3$, with $a_c=0.5$, $a_M=0.00$, $K=400$, $N=1\,000$, and $Le=16$

(a) First trial (b) Second trial (c) Third trial

best-adapted species (Group A) before the 100th generation.

A population increase is not observed for Group B before the best-adapted species emerges widely. This small population of Group B in the environment of quasi-macroevolution was also seen in the previous report⁽⁸⁾.

4. Computational Experiments for Various Crossover and Mutation Rates

Let us consider why long introns accelerate the quasi-macroevolution described in the foregoing section. Attention is paid to the fact that, while the overall crossover rate a_c is fixed in Section 3, the

Fig. 3 Time-dependent appearance probabilities in the environment for quasi-macroevolution of $Lc=3$ with $a_c=0.5$, $a_M=0.00$, $K=400$, $N=1\,000$, and $Le=16$

(a) $L=16$ ($Li=0$) (b) $L=32$ ($Li=16$)
(c) $L=64$ ($Li=48$) (d) $L=256$ ($Li=240$)

possibility of crossover occurring in exons decreases as the number of introns increases. As expected, the SGAs of smaller overall crossover rates without introns also accomplish quasi-macroevolution at $L_c=3.0$ (see Fig. 4).

An answer to the question of why the rate of crossover occurring in exons must be reduced for quasi-macroevolution to occur is basically shown in the analysis of microevolution in the previous reports^{(7),(8)}. Even in microevolution, extremely excessive crossover rates can not assist evolution^{(7),(8)}. The region of the crossover rate at which evolution is realized is extremely narrow for quasi-macroevolution in comparison with microevolution. Thus, in order to set the crossover rate in that narrow region of crossover rate, a tolerable reduction of the crossover rate is necessary for quasi-macroevolution. However, it should be stressed that the complete elimination of the crossover operator leads to the instability of evolution^{(7),(8)}.

The following observations also indicate that a function of introns is to reduce the rate of crossover occurring in exons.

(1) Computational experiments of 1 502 trials for $Li < 485$ were also performed for two other crossover

Fig. 4 Time-dependent appearance probabilities in the environment for quasi-macroevolution of $L_c=3$ with $\alpha_M=0.00$, $K=400$, $N=1\,000$, $Le=16$, and $Li=0$
(a) $\alpha_c=0.02$ (b) $\alpha_c=0.10$

rates. Figure 5 shows the shortest intron length (the critical intron length) at which the best-adapted species emerges widely. (Since three trials for $0 < Li < 485$ were performed at each crossover rate, three dots are shown in Fig. 5.) In the region over the critical line shown in Fig. 5, the best-adapted species emerges widely, although the best-adapted species can not emerge in the region below the critical line, i.e., the correlative line for the dots shown in Fig. 5. The relation between the overall crossover rate α_c and the critical value of intron length Li at which quasi-macroevolution occurs is linear.

(2) Figure 6 shows that mutation has less influence on the critical value of intron length Li at which such evolution occurs.

(3) Attention is also paid to the fact that systems with small crossover rates prefer short introns, as is observed in Figs. 5 and 6. Actually, yeast such as *Saccharomyces cerevisiae* is well known as a short intron system with a small crossover rate⁽¹⁰⁾.

Fig. 5 The relation between the crossover rate α_c and the critical value of L at which quasi-macroevolution of $L_c=3$ occurs with $\alpha_M=0.00$, $K=400$, $N=1\,000$, and $Le=16$

Fig. 6 The relation between the crossover rate α_c and the critical value of L at which quasi-macroevolution of $L_c=3$, occurs with $\alpha_M=0.05$, $K=400$, $N=1\,000$, and $Le=16$

(4) The string length increase due to the addition of introns can not change the environment, although the string length increase due to the duplication of exons can vary the environment essentially, as is discussed in the previous report⁽⁸⁾.

(5) We have also applied SGAs without crossover, that is, SGAs only with replication and mutation. In such cases, the best-adapted species did not emerge in either the intron-exon system or the system without introns. This may be explained by the characteristic that crossover directly generates Group A (the best-adapted species) from Group C, although one-point mutation is impossible^{(7),(8)}.

5. Computational Experiments for Various Replication Rates

Figures 5 and 7 show the influence on the critical intron length of the variation in the threshold value K at $L_c=3.0$. Here, let us pay attention to the critical ratio Li/Le for $a_c=1.0$, that is, for the full crossover system. The critical intron length in the full crossover system increases as the threshold value increases, that is, as the replication rate increases. It should be noted that, in fact, higher Eukaryotes with high replication and crossover rates have long intron systems⁽¹⁰⁾.

6. Computational Experiments for Various Environments

The relation between the overall crossover rate and the critical value of Li at which quasi-macroevolution occurs was obtained by the computational experiments for various lengths of L_c (see Fig. 8). Quasi-macroevolution can be accomplished by using long introns even in the environments where the structure of the best-adapted species is farther from those of the other types of genes.

Figure 8 shows that the critical intron length increases with increasing L_c . It is emphasized that the critical ratio of Li/Le is often over 10.0, as is observed in actual higher Eukaryotes⁽¹⁰⁾.

Microevolution was stably accomplished for a wide range of overall crossover rates⁽⁶⁾. This implies that the overall crossover rates need not be optimized for microevolution very much, and it suggests that introns will not accelerate microevolution.

7. Conclusions and Discussion

7.1 Conclusions

By developing an artificial brain, which incorporates a learning theory such as the genetic algorithm (GA), researchers have tried to build autonomic control systems, such as those for automobiles. However, artificial brains are still at the level of a newborn baby, who can not judge anything and can not drive a

car. The problem of safety and security is a critical issue for automobiles. At a minimum, we need some method for optimizing the parameters in the GA according to the environment.

In the previous reports^{(7),(8)}, methods were proposed for clarifying the essential differences between problems (the UDS and the IDS) and for inferring the characteristics of problems with and

Fig. 7 The relation between the crossover rate a_c and the critical value of L at which quasi-macroevolution of $L_c=3$ occurs with $q_M=0.00$, $K=300$, $N=1\,000$, and $Le=16$

Fig. 8 The relation between the crossover rate a_c and the critical value of L at which quasi-macroevolution occurs with $a_M=0.00$, $K=400$, $N=1\,000$, and $Le=16$ (a) $L_c=4$ (b) $L_c=5$

without introns. An intuitive judgment model was proposed for optimizing the parameters in cases where introns do not exist. However, the model can not be used for cases with introns.

This report has proposed a method for optimizing the crossover rate in cases where introns exist. Artificial macroevolution can be realized by reducing the overall crossover rate or by increasing the ratio of intron and exon lengths for deceptive problems.

7. 2 Discussion

(a) Macroscopic Kinetic Equation for the SGA with Introns

The macroscopic kinetic equation developed previously⁽⁸⁾ can also predict the evolutionary routes in the SGA with introns, by using the crossover and mutation coefficients⁽⁸⁾ for the exon length Le , by reducing the overall crossover rate inversely proportional to the intron length and by evaluating the environment from the genes of string length Le which have eliminated introns.

(b) Position of Introns

In the present report, we treated a simple system with one intron domain and one exon domain. The locations of introns do not influence the present results, since the positions of introns have no roles in the three genetic operators of crossover, mutation, and replication.

(c) Evolution Machines with Continuous Homologous Recombination Technique

In the future, evolution machines⁽¹²⁾ will be improved by including the artificial recombination technique used in genetic engineering. The homologous recombination explained in molecular biology is the crossover operator used in the GA. Therefore, artificial macroevolution (new species) may be obtained when the rate of artificial recombination used in continuous bioreactors is optimized on the basis of the knowledge obtained in the present study.

(d) Function of Introns in Actual Eukaryotes

The population size is assumed to be constant. It is said that the number of various species is adjusted manually to maintain a constant overall population level in laboratory experiments and that a constant population system was present on the Earth in ancient times⁽⁴⁾. Evolution occurs within the Earth's closed system with a limited amount of resources. Thus, the assumption of a constant population size is not exceptional.

The strongly nonlinear replication operator used in the present study is observed in actual life systems⁽⁴⁾.

In the problem shown in Section 2, the frequency distribution plotted against the fitness value for all the possible types of genes has a bi-modal distribution,

since the probability is separated by the total weight limitation represented in Eq. (2.b). This frequency distribution is observed in actual life systems such as the *Drosophila pseudoobscura*⁽¹³⁾.

Thus, analyses based on the SGA will lead to a better understanding of actual Eukaryotes.

Levenick⁽⁹⁾ examined the relation between the mutation rate and introns in the environment of microevolution. The present work has shown the relation between the crossover operation and introns in the environment of quasi-macroevolution. The results of the present analyses indicate that systems with high crossover and replication rates often prefer long introns of $Li/Le > 10.0$, as is observed in actual higher Eukaryotes, and also that systems with small crossover rates have relatively short introns, as is seen in actual yeasts. One function of introns in actual Eukaryotes is to accelerate quasi-macroevolution. The correlative relation between introns and mutation is relatively weak.

Acknowledgments

The author is deeply indebted to Professor Egon Krause of the Aerodynamisches Institut der RWTH Aachen for his encouragement on this study. This study was partially conducted at the 2nd and 3rd workshops on "Various Approaches to Complex Systems" held at the International Institute for Advanced Studies in Kyoto in 1995-1996.

References

- (1) Goldberg, D.E., Genetic Algorithms, (1989), Addison-Wesley Publishing Company Inc., Reading, Massachusetts.
- (2) Kitano, H., Combination of Genetic Algorithm and Neural Network, in Genetic Algorithm, ed. Kitano, H., (1993), Sangyo Tosho, Tokyo.
- (3) Edelman, G.M., Neural Darwinism, (1989), Oxford University Press, Oxford.
- (4) Eigen, M., Steps towards Life, (1989), Oxford University Press.
- (5) Rechenberg, I., Evolutionalstrategy, (1973), Frommann-Holzboog Verlag, Stuttgart.
- (6) Vose, M., Punctuated Equilibria in Genetic Search, Complex Systems, Vol. 5 (1991), p. 31-44.
- (7) Naitoh, K., Four-Group Equation of Genetic Algorithm, JSME Int. J., Ser. C, Vol. 38, No. 2 (1995), p. 240.
- (8) Naitoh, K., Macroscopic Kinetic Equation for a Genetic Algorithm, Japan Journal of Industrial and Applied Mathematics, Vol. 15, No. 1 (1998). (Also in Bussei Kenkyu (Kyoto), Vol. 66, No. 5 (1996), p. 896).
- (9) Levenick, J.R., Inserting Introns Improves Genetic Algorithms Success Rate: Taking a Cue from Biology, ICGA-4: Proceedings of the Fourth

- International Conference on Genetic Algorithms, (1991), p. 123-127, Morgan Kaufmann
- (10) Watson, J.D., Hopkins, N.H., Roberts, J.W., Steitz, J.A. and Weiner, A. M., Molecular Biology of the Gene, Fourth Edition, (1987), The Benjamin/Cummings Publishing Company.
- (11) Konagaya, A., Genetic Algorithm for Combination Problems, in Genetic Algorithm, ed. Kitano, H., (1993), Sangyo Tosho, Tokyo.
- (12) Husimi, Y. and Keweloh, H.C., Continuous Culture of Bacteriophage Qb Using a Cellstat with a Bubble-wall Growth Scraper, Rev. Sci. Instrum. Vol. 58 (1987), p. 1109.
- (13) Smith, J.M., Evolutionary Genetics, (1992), Oxford University Press, Oxford.
-