DNA ALGORITHMS BASED ON EXON SHUFFLING

Tsuyoshi Okayama and Haruhiko Murase

Osaka Prefecture University
1-1, Gakuen, Sakai 599-8531 Japan

Abstract: An understanding of a natural system's information handling can lead to more effective artificial optimization techniques. There are successful optimization algorithms represented in biosystems that have proven useful in engineering applications (artificial neural networks, immune system algorithms, etc). The goal of our study is to develop a new biosystem derived an optimization algorithm which is called a DNA algorithm (DNAA) based on optimization procedures in DNA. We have focused on an analogy between optimizing procedures for protein functions using exon shuffling and those for an optimization problem in the engineering field. We used a traveling salesman problem (TSP) for evaluation of the performance of the DNAA. The DNAA could estimate approximately optimal tour routes in the 25-city TSP. *Copyright © 2005 IFAC*

Keywords: intron, exon, exon shuffling, optimization algorithm, traveling salesman problem.

1. INTRODUCTION

An understanding of a natural system's information handling can lead to more effective artificial optimization techniques. A genetic algorithm (GA) is one of the successful instances. GAs are robust optimization procedures which have been widely utilized in areas such as engineering optimization and design, forecasting, image recognition and functional optimization, because an optimal value can be searched for in parallel with a multi-point search technique, rather than a single point procedure (Goldberg, 1989; Holland, 1992). Even if the objective function has many peaks, the multipoint search technique permits the focus of attention on the most valuable parts of the solution space and consequently, the global optimal value can be rapidly and efficiently to guide its search. There is no requirement to formulate a mathematical equation or any a priori knowledge for the objective function. There are more other algorithms represented in biosystems that have proven useful in engineering applications (artificial neural networks, immune system algorithms, etc). We have also developed photosynthetic algorithms based on photosynthetic reactions (Okayama and Murase, 2002a) and leaf cellular automata based on the photosynthetic activities on and in leaves (Okayama and Murase, 2002b). The goal of our study is to develop a new biosystem derived an optimization algorithm which is called a DNA algorithm (DNAA) based on optimization procedures in DNA. Using introns (I will explain what 'intron' is later) has been previously applied to genetic algorithms by Levenick (1991). His concept could improve the success ratio of evolution of artificial flying creatures, because the insertion of introns increases the number of crossover points that preserve genes. Surprisingly, we can find the same concept in biological theory that was reported Fedorova and Fedorov (2003). The result of Levenick's paper indicates the possibility of the DNA derived algorithm. In addition the facts of DNA that are useful as optimization techniques have been revealed, because in the last couple of decades enormous number of researches concerning DNA has been conducted. DNA is the genetic material that is propagated from generation to generation, and contains the instructions on how to build the proteins necessary for a particular organism. Though all genetic information is stored in the ordering of the nucleotides in the DNA, DNA is not directly involved in protein synthesis. DNA directs protein

synthesis by sending instructions in the form of RNA. RNA carries out the synthesis of proteins from the DNA instructions. Introns are segments of DNA found within gene. On the contrary, exons are segments that have information about the synthesis of proteins. Introns are transcribed into RNA along with the rest of the gene but must be removed from the RNA before the mature RNA product is complete. RNA that still contains the intron regions is often called pre-RNA. After the introns are spliced out of the pre-RNA, the remaining segments of RNA, the exons are joined together to become the mature RNA product. Introns are present in all studied eukaryotic organisms. Relatively small number of introns was found in single-cell organisms while dozens of thousands were discovered in the completely characterized genomes of plants, invertebrates, and vertebrates (Fedorova and Fedorov, 2003). Evolution of intron-exon structure is still controversial. What is the role of introns in the genome? Notwithstanding that introns were discovered 25 years ago, there are still opposite viewpoints on this question. Fedorova and Fedorov (2003) reviewed six distinct roles of introns: (1) sources of non-coding RNA; (2) carriers of transcription regulatory elements; (3) actors in alternative and trans-splicing; (4) enhancers of meiotic crossing over within coding sequences; (5) substrates for exons shuffling; and (6) signals for mRNA export from the nucleus and nonsense-mediated decay. We have focused on the (5) substrates for exon shuffling as an optimization technique of the DNAA.

Specific objects of this study are 1) to establish a concept of the DNA algorithm, and 2) to investigate the performance of the DNAA using a traveling salesman problem.

2. PEVIEW OF THE DNA ALGORITHM

2.1 Introns and Exion Shuffling.

We have focused on the function of introns and exon shuffling as an optimization technique of the DNAA. The intron-exon organization of eukaryotic genes suggests that new combinations of exons can be created by recombination within the intervening intron sequences, yielding rearranged genes with altered functions. This evolutionary mechanism of recombining exons from unrelated genes is known as exon shuffling. It has been proposed that exon shuffling has been a major factor in protein evolution (Kolman and Stemmer, 2001). Introns-early theory of the assembly of genes from exon ‘pieces’ was elaborated in details as ‘the exon theory of genes’ by Gilbert (1987). In the original form this theory describes to the earliest steps of evolution starting at the time of ‘RNA world’. The theory claims that ancient genes were assembled from mini-exons during the exon-shuffling process. It proposed a simple scheme of molecular events at the genomic level that could dramatically increase the speed of gene evolution. Under this scheme, instead of trying practically unlimited numbers of gene mutations to produce functional proteins, genes were assembled from exonic ‘pieces’ coding functional protein segments (Patthy, 1996). Fig. 1 shows an instance of exon shuffling. In this study we have focused on exon shuffling. We have focused on an analogy between optimizing procedures for protein functions and those for an optimization problem in the engineering field. In other words, we consider exons as ‘building block’ which can be a part of high fitness solution, exon shuffling can be used as an optimization technique.

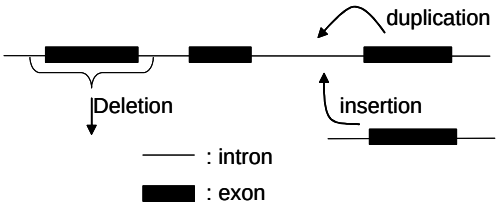

Fig. 1. Exon shuffling

2.2 Flow chart of DNAA

Fig. 2 shows flowchart of DNAAs. First initial DNAs are generated randomly. Then, introns of which lengths are unity are inserted randomly with a probability of an intron insertion ratio. If an intron insertion point is adjoined or in an intron, the length of the intron is lengthened. On the contrary, if an intron is inserted into an exon, the exon is divided into two exons. Then exon shuffling is executed. Usually, exon shuffling could contain some actions such as, exchange, insertion, deletion, and duplication of exons. It depends on the optimization problem which procedures of exon shuffling are suitable for. Then, all of DNA are spliced and become RNA to be decoded. In order to search an optimized solution efficiently, the tournament selection strategy is used for improving average fitness of the population.

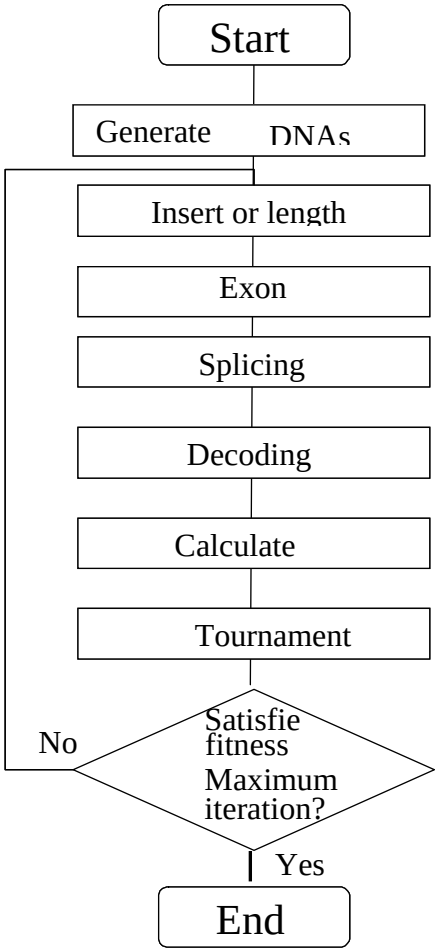

Fig. 2. Flow diagram.

Tournament selection begins by randomly selecting *n* candidates from the population. The candidates in the tournament selection are then pitted against each other by comparing their fitness. If there is a single candidate with the highest fitness, this candidate is the winner of the tournament and is remained in the next generation. If two or more candidates have the same fitness, the winner is selected randomly among them. These procedures are continued until the number of population reaches the target value. Then, some introns are removed from DNAAs randomly with a probability of an intron deletion ratio. These procedures are continued until the best fitness of the population reaches target fitness, or a number of iteration exceeds a target number.

3. TRAVELING SALESMAN PROBLEMS

We used a travelling salesman problem (TSP) for evaluation of the performance of the DNAA. The TSP is one of the most widely discussed problems in combinational optimization. In the TSP, hypothetical salesman must make a complete tour of a given set of cities in the order that minimizes his total tour distance. Fig.3 shows that one of RNA represents a salesman's visiting order. To make sure that the salesman visits only once in his tour, the DNAA employed only exchange procedure for the exon shuffling. In the experiment, we use the 25-sity TSP. Intron insertion ratios were varied from 0.01 to 0.10 at 0.01 intervals and intron deletion ratios were varied from 0.05 to 0.50 at 0.05 intervals. To avoid waste of computer memory, the limit of total length of introns were set 1000.

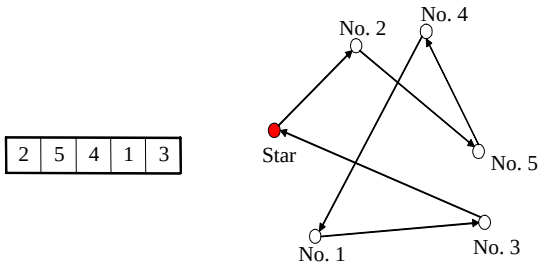

Fig.3. RNA and the route.

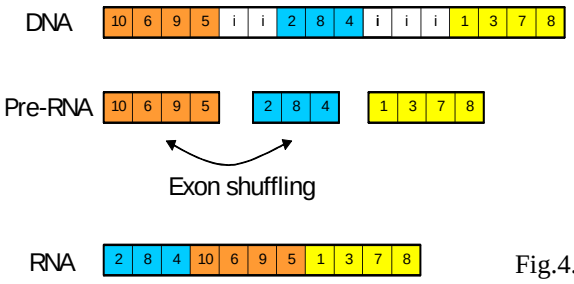

Fig.4. Splicling and Exson Shuffling.

4. RESULTS AND DISCUSSIONS

Fig.5 shows a typical initial route (left) and a successful route (right) after 1000 iteration. Although the route in Fig.5 is not perfectly optimized, the route is nearly shortest. Fig.6 shows tour lengths of each combination of intron insertion ratios and intron deletion ratios. When intron insertion ratio was below 0.02, the tour lengths were always over 600. When the intron insertion ratio was over 0.07 and the intron deletion ratio was under 0.2, tour lengths were usually over 550. When the ratio of the intron insertion ratio to the intron deletion ratio was 1 to 5, most tour lengths were under 525. In order to discuss simply, we picked up typical combinations of intron insertion ratios and intron deletion ratios and classified into three groups. Table 1 shows the combinations included in these groups. The group 1 includes the combinations of low intron insertion ratios and relatively high intron deletion ratios. Their tour lengths are over 625. The group 2 includes the combinations of high intron insertion ratios and low intron deletion ratios. Their tour lengths are over 600. The group 3 includes the combinations which ratios of intron insertion ratios to intron deletion ratios are approximately 1:5. Their tour lengths are under 525.

Namely speaking, Group 1 and 2 include typical combinations of intron insertion ratios and intron deletion ratios have led inefficient tours, and group 1consists of the successful combination of them. Fig.7 shows changes of tour lengths in each group. The tour length in group 1 was decreased more slowly than those of other two groups. The length in the group 2 decreased rapidly until about 200 iterations, but the length was hardly changed after 200 iterations. In group 3, the length was decreased rapidly until about 300, and it was still decreased slowly. Fig.8 shows change of a number of introns in each DNA. The number of introns in the group 1 was always around 0. The number of introns in the group 2 was increasing until 200 iterations and reached over 12. Then the number was decreasing slowly. This is because the maximum number of introns in each DNA was set 1000. The number of introns in the group 3 was varying between 1 and 2. These results indicate that keeping two or three building blocks are suitable for the 25-city TSP problem. These results indicate that the low intron insertion ratios disturb rapid decreasing of the tour length and the high intron insertion ratios contributed to rapid decrease of the tour length. The result of the group3 indicated that the ratio of the intron insertion ratio to the intron deletion ratio that was suitable for this TSP was 1:5.

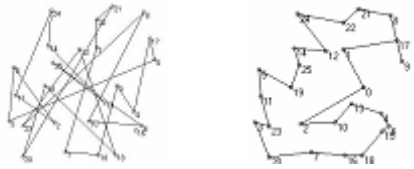

Fig. 5 An initial route (left) and an instance estimated route after 1000 iterations (right).

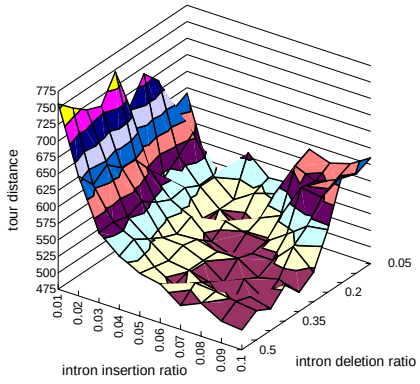

Fig.6. Tour length after 1000 iteration

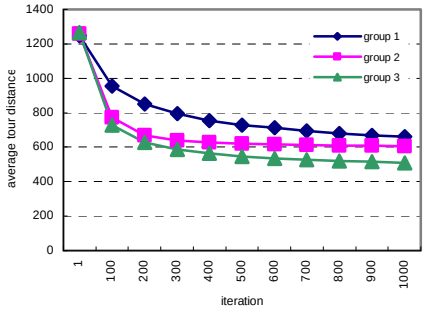

Fig. 7. Change of average tour length

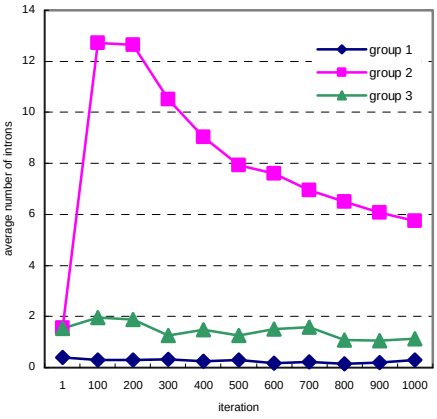

Fig. 8. Change of average number of introns in each DNA

5. CONCLUSION

In this study, DNAA has been developed based on ‘exon shuffling’, and DNAA could estimate approximately optimal tour routes in the 25-city TSP. We used one exchange as a procedure in exon shuffling, DNAA could employ other procedures in exon shuffling, such as insertion, deletion, and duplication for other types of optimization problems. Of course, DNAA can be combined with GA approaches, such as crossover, mutation, etc. And also a lot of optimization system of DNA can be harnessed for the biosystem derived algorithm. There are surely many other algorithms represented in biosystems that might prove useful in engineering applications. Seeking useful engineering principles exemplified in biosystems is likely to be a fruitful path to advancements in bioengineering.

Table 1 Combinations of intron insertion ratio and intron deletion ratio in each group

group 1	group 2	group 3
(0.01, 0.35)	(0.07, 0.05)	(0.07, 0.35)
(0.01, 0.40)	(0.07, 010)	(0.08, 0.30)
(0.01, 0.45)	(0.08, 0.05)	(0.08, 0.35)
(0.01, 0.50)	(0.08, 0.10)	(0.08, 0.40)
(0.02, 0.40)	(0.09, 0.05)	(0.09, 0.35)
(0.02, 0.45)	(0.09, 0.10)	(0.09, 0.40)
(0.02, 0.50)	(0.10, 0.05)	(0.09, 0.45)
(0.03, 0.40)	(0.10, 0.10)	(0.10, 0.40)
(0.03, 0.45)	(0.10, 0.15)	(0.10, 0.45)
(0.03, 0.50)		(0.10, 0.50)

(intron insertion ratio, intron deletion ratio)

REFERENCES

Fedorova, L., A. Fedorov (2003). Intron in gene evolution. *Genetica* **118**, 123-131.

Gilbert, W. 1986. The RNA world. *Nature*, **319**, 618.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA: Addison Wesley.

Holland, J.H. (1992). Genetic algorithms. *Sci. Am.*, July: 44-50.

Kolman, J.A., and W. P.C. Stemmer (2001). Direct evolution of proteins by exon shuffling. *Nature* Publishing Group http://biotech.nature.com

Okayama, T., H. Murase (2002). Leaf Cellular Automata, *Japanese Society of high technology in agriculture*, **14** (3), 34-38.

Okayama, T., H. Murase (2002). Solution for N queens Problem Using A Photosynthetic Algorithm. *Proceedings of the 15th IFAC World Congress 2002*, Vol.T, 335-338.

Patty, L. (1996) Exon Shuffling and Other Ways of Module Exchange. *Matrix Biology* **15**, 301-310.

Levenick, J.R. (1991). Insertion Introns Improves Genetic Algorithm Success Rate: Taking a Cue from Biology. *Proceedings of the Fourth International Conference on Genetic Algorithmds*. San Mateo California: Morgan Kaufmann, 123-127.

Wu, A.S., and K. Lindsay (1996). A survey of intron research in genetics. *Proceedings of the 4th Conference on Parallel Problem Solving from Nature* Berlin, Germany, September 1996, 101-110.