DNA Algorithm and Its Application

Tsuyoshi Okayama^{*1)} and Haruhiko Murase²⁾

¹⁾ Tokyo University of Agriculture and Technology, Saiwai 3-5-8, Fuchu, 183-8509, Japan

²⁾ Osaka Prefecture University, Gakuen-cho 1-1, Sakai, 599-8531 Japan

Abstract

An understanding of the way a natural system handles information can lead to more effective artificial optimization techniques. There are successful optimization algorithms found in biosystems that have proven useful in engineering applications (artificial neural networks, genetic algorithms, immune system algorithms, etc). Okayama *et al.* developed a new biosystem-derived optimization algorithm called a DNA algorithm (DNAA). They found that the insertion of introns is effective for protecting important exons from mutation and DNAA's performed robustly even under a fairly high rate of mutation in string search problems and knapsack problems. In this study, we have focused on exon shuffling as a new optimization technique for the DNAA. The intron-exon organization of eukaryotic genes suggests that new combinations of exons can be created by recombining them 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. We consider modules as 'building blocks' which can be a part of high fitness solutions. To evaluate the performance of the DNAA we used a traveling salesman problem (TSP), one of the most widely used challenges for combinatorial optimization. In the experiment, we used a 100-city TSP. The intron insertion ratio and intron deletion ratio were varied from 2 to 128. The DNAA could approximate optimal tour routes in the 100-city TSP.

Keywords

intron, exon, exon shuffling, optimization algorithm, traveling salesman problem

Introduction

An understanding of the way a natural system handles information can lead to more effective artificial optimization techniques. One successful example is genetic algorithms (GAs). 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).

There are several other algorithms represented in biosystems that have proven useful in engineering applications, such as artificial neural networks, immune

system algorithms, etc. Okayama and Murase developed algorithms based on photosynthetic reactions (2002a) and for leaf cellular automata based on the photosynthetic activities on and in leaves (2002b).

In a previous study, Okayama *et al.* (2003) developed a new biosystem-derived optimization algorithm which is called a DNA algorithm (DNAA). In the present study, it was found that the insertion of introns is effective for protecting important exons from mutation, and DNAA's performed robustly even under a fairly high rate of mutation in string search problems and knapsack problems. In the knapsack problems, it was found that more introns were accumulated near exons whose role seemed more important than other exons.

Previously, introns had been applied to genetic algorithms by Levenick (1991). His concept could

* Corresponding Author

E-mail: tsu-okay@cc.tuat.ac.jp

improve the success rate of the evolution of artificial flying creatures, because the insertion of introns increases the number of crossover points that preserve genes. Surprisingly, the same concept can be found in biological theory that was reported by Fedorova and Fedorov (2003). These results indicate the possibility of finding the mechanism of DNA-derived algorithms. In addition, DNA data that are useful for optimization techniques have been obtained in the last couple of decades as a result of the enormous number of DNA studies that have been conducted.

The mechanism of the DNA has a lot of interesting features. Introns are segments of DNA found within a gene, while 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 completed. 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. A relatively small number of introns have been found in single-cell organisms while tens of thousands have been discovered in the completely characterized genomes of plants, invertebrates, and vertebrates (Fedorova and Fedorov 2003).

Despite the fact that introns were discovered 25 years ago, there is still much left to be learned about the evolution of the intron-exon structure. 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. All these six assumption could be utilized for optimization techniques for DNAAs.

In this study, we have focused on (5), the substrates for exon shuffling, as a new optimization technique for DNAA. The intron-exon organization of eukaryotic genes suggests that new combinations of

exons can be created by recombining them 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. Exon shuffling may be a major factor in protein evolution (Kolkman and Stemmer 2001).

The introns-early theory of the assembly of genes from exon ‘pieces’ was elaborated in detail as ‘the exon theory of genes’ by Gilbert (1987). In its original form, this theory describes the earliest steps of evolution starting at the time of the ‘RNA world’. The theory claims that ancient genes were assembled from mini-exons during the exon-shuffling process. It proposes 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 the present study, 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 modules as ‘building blocks’ which can be a part of high fitness solutions.

Specific objectives of this study are 1) to improve DNA algorithms using the exon shuffling concept as a new additional optimization technique and 2) to investigate the performance of the DNAA using a traveling salesman problem.

Review of DNA Algorithms with Exon Shuffling

Fig. 2 shows a flowchart of DNAAs. First, an initial DNA is generated randomly. The DNA consists

Fig. 1 Sample of exon shuffling.

of exons whose lengths are decided arbitrarily and introns are located between exons. Next, introns are selected randomly and deleted with a probability of an intron deletion ratio. Then two exons that are neighbors of the intron are combined into one exon. In contrast, an intron is inserted into an exon, which is then divided into two exons with a probability of an intron insertion ratio. Then, exon shuffling is executed. Usually, exon shuffling can involve such actions as exchange, insertion, deletion, and duplication of exons, depending on the optimization problem. Then, the DNA is spliced and becomes an RNA, which is evaluated as a solution. If the solution is better than the previous fitness, the RNA is transcribed into the DNA. If not, the RNA is discarded. These procedures are continued until a solution is obtained, or a number of iterations reaches a limited number.

Fig. 2 Flowchart of the DNAA.

Traveling Salesman Problem and Coding

To evaluate the performance of the DNAA, a traveling salesman problem (TSP), one of the most widely used techniques in combinational optimization, was employed. In the TSP, a hypothetical salesman must make a complete tour of a given set of cities in the order that minimizes his total tour distance. In this experiment, cities were located in a 100 X 100 square area. Fig.3 shows that an RNA represents a salesman's visiting order.

The intron deletion ratio

The shorter the distance between exons, the easier it is to delete an intron between the exons. The

Fig. 3 An RNA represents a salesman's route.

intron deletion ratio is calculated using the following equation:

Intron deletion ratio = D_v / Distance between exons (1)

Where D_v is an arbitrary value that should be decided manually. D_v was varied from 2 to 128.

The intron insertion ratio

The longer the distance between bases (bases indicate A, T, G and C in DNA, in this problem bases represent cities), the easier it is to insert an intron between bases. The intron insertion ratio is calculated using the following equation:

Intron insertion ratio = Distance between bases / I_v (2)

Where I_v is an arbitrary value that should be decided manually. I_v was varied from 2 to 128.

In order to make sure that the salesman visits each place in his tour only once, the DNAA only employed an exchange procedure, the 100-city TSP, for the exon shuffling (Fig. 4).

Simple genetic algorithm

In order to evaluate the performance of the DNAA objectively, a simple genetic algorithm

Fig. 4 Splicing and exon shuffling in the DNAA.

(GA) was employed. The population size, mutation ratio and the way of crossover of the simple GA was 10 were 10, 0.05 and one-point crossover, respectively. The elite selection strategy and tournament selection strategy were employed to generate next generation (Goldberg 1989).

RESULTS AND DISCUSSION

Fig.5 shows a typical initial route (left) and a successful route (right) after 100,000 iterations. Although the successful route in Fig.5 is not perfectly optimized, the route is nearly the shortest.

Fig.6 shows the shortest tour lengths which were estimated after 100,000 iterations with each combination of D_v and I_v . The data were averaged over 10 runs. The combination of $D_v=8$ and $I_v=2$ could find the shortest route ($=789.55$). The combination of $D_v=128$ and $I_v=32$ shows the longest route ($=1246.11$). When the product D_v and I_v was 16, 32 or 64, most of estimated routes were less than 850. This indicates that the balance between D_v and I_v was an important factor for estimation.

Fig. 7 shows the estimated tour length after 100,000-iterations and Fig. 8 shows the number of exons during the estimation of the combinations $D_v=8$ and $I_v=2$, $D_v=4$ and $I_v=1$, and $D_v=128$ and $I_v=32$, respectively. The number of exons of the combination of $D_v=8$ and $I_v=2$, which could find the shortest route, was varied around between 20 and 40. In contrast, the number of exons in the combination of $D_v=128$ and $I_v=32$ which found the longest route was always 3 or 4. This means this combination could not search shorter routes globally, so the result was worse.

The number of exons in the combination of $D_v=4$ and $I_v=1$ was varied between 40 and 60. Because of the excessive number of exons, this combination could not search the shorter routes locally. These results indicate the balance between global search and local search was important for the optimization. They also showed that the DNAA could control the balance using variation of combination of D_v and I_v .

Fig. 9 shows the results when the number of exons was maintained at 30. The convergence speed was much slower than that of the combination of $D_v=8$

and $I_v=2$. These results indicate that local optimization using Eq. 1 and Eq. 2 contributed to the quick convergence in the early stage of the search.

Fig. 10 shows the result of the tour lengths of the DNAA which had the combination of $D_v=8$ and $I_v=2$ and the simple GA. It is very difficult to compare the performance of different algorithms, because the DNAA used one solution in each iteration, and the simple GA used 10 solutions (chromosomes) in each generation. Therefore, this comparison of the performance between the DNAA and GA was not proper comparison. However, this result shows that the DNAA might be a better algorithm than the simple GA. This is because the DNAA could optimize locally more efficiently than the simple GA using Eq. 1 and Eq. 2.

CONCLUSION

In this study, a DNAA has been developed based on 'exon shuffling', and that DNAA could approximate optimal tour routes in the 100-city TSP. While we used an exchange procedure in exon

Fig. 5 An initial route (left) and an instance estimated route after 100,000 iterations (right).

Fig. 6 Tour length after 100,000 iteration.

Fig. 7 Change of average tour length.

Fig. 8 Change of number of exons.

Fig. 9 Change of average tour length between uncontrolled and controlled number of exons.

Fig. 10 Comparison of calculated tour lengths of the DNAA and the simple GA.

shuffling, DNAA could employ other procedures in exon shuffling, such as insertion, deletion, and duplication for other types of optimization problems. In addition, a lot of optimization systems of DNA could potentially be harnessed for biosystem-derived algorithms. There are surely many other algorithms represented in biosystems that might prove useful in engineering applications. The search for useful engineering principles exemplified in biosystems is likely to contribute to advancements in bioengineering.

References

- Fedorova, L., and 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.*, 267: 66-72.
- Kolkman, J.A. and W. P.C. Stemmer (2001) Direct evolution of proteins by exon shuffling, *Nature Publishing Group* <http://biotech.nature.com> browsed on Nov. 5, 2003.
- Levenick, J.R. (1991) Introns Improves Genetic Algorithm Success Rate: Taking a Cue from Biology, In Belew, Richard K., Lashon B. Booker (editors), *Proceedings of the Fourth International Conference on Genetic Algorithms*, San Mateo California: Morgan Kaufmann: 123-127.
- Okayama, T., H. Kitabata and H. Murase (2003) DNA Algorithm, *CJapanese Society of Agricultural Informatics*, 12 (1): 33-44.
- Okayama, T. and H. Murase (2002a) Solution for N queens Problem Using A Photosynthetic Algorithm, *Proceedings of the 15th IFAC World Congress 2002*, Vol.T, 335-338.
- Okayama, T. and H. Murase (2002b) Leaf Cellular Automata, *J. Soc. of High Tech. in Agr.*, No.14 (3): 152-156.
- Patty, L. (1996) Exon Shuffling and Other Ways of Module Exchange, *Matrix Biology*, 15: 301-310.

Received October 08, 2004

Accepted June 27, 2005

Information engineering.

DNA アルゴリズムとその応用

岡山 毅^{1)*}, 村瀬治比古²⁾

¹⁾東京農工大学 農学部

²⁾大阪府立大学 農学生命科学研究科

要旨

自然界の情報処理系には、進化の結果として得られた数多くの最適化手法が確認されている。これらの最適化手法の理解は、新しい最適化アルゴリズムの構築につながる。その成功例として、遺伝的アルゴリズム、ニューラルネットワーク及び免疫的アルゴリズム等が挙げられる。岡山らによって提案された、それらに続く新しい生物系由来最適化アルゴリズムである DNA アルゴリズムは、遺伝子情報を含まないとされるイントロンを利用しており、イントロンをエキソン部に挿入することにより、重要な遺伝子情報を含むエキソン部を突然変異から保護し、結果として突然変異率の設定に対して、安定した解の推定が可能である。本研究では、この DNA アルゴリズムに、新たにエキソンシャッフリングの概念を導入し、その効果を検証するために 100 都市巡回セールスマン問題を用いた。その結果、イントロン挿入率およびイントロン削除率を変化させ、エキソン数を調整することで、ほぼ最短の巡回ルートの探索が可能であった。また、単純遺伝的アルゴリズムと比較したところ、10 個体、10,000 世代において、単純遺伝的アルゴリズムの推定ルート長が 1997 に対し、DNA アルゴリズムでは 881 と良好な結果が得られた。

キーワード

イントロン, エキソン, エキソンシャッフリング, 最適化アルゴリズム, 巡回セールスマン問題

* Corresponding Author : E-mail: tsu-okay@cc.tuat.ac.jp