2010 International Symposium on Nonlinear Theory and its Applications
NOLTA2010, Krakow, Poland, September 5-8, 2010

Heuristics Methods for Asymmetric Traveling Salesman Problem and their Applications to DNA Fragment Assembly

Tomohiro Kato† and Mikio Hasegawa†

†Dept. of Electrical Engineering, Tokyo University of Science
1-14-6 Kudankita, Chiyoda-ku, Tokyo, 102-0073 Japan

Abstract—There are various applications of the Traveling Salesman Problems(TSPs) in a real world. In this paper, a DNA fragment assembly problem is studied as one of the applications of the TSP. The DNA fragment assembly problem is to build a DNA sequence from several hundreds of fragments obtained by the genome sequencer, which can be formulated as the asymmetric TSP. We apply heuristic methods for the asymmetric TSP, such as the tabu search, the simulated annealing, and the genetic algorithm, to the DNA fragment assembly problem. Our simulation results show that the proposed algorithm using the tabu search on a block shift operation exhibits the best performance for the asymmetric TSPs and the DNA fragment assembly problems.

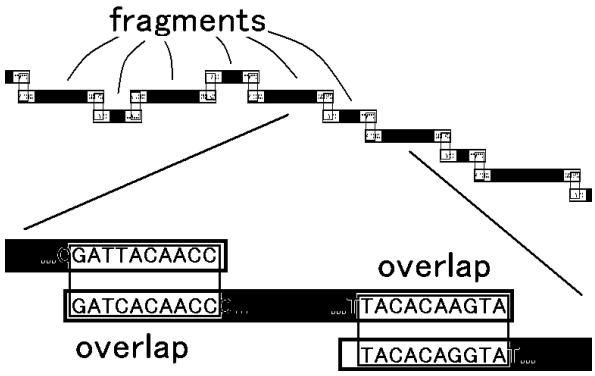

Figure 1: DNA fragment assembly problem

1. Introduction

In recent years, genetic technologies are progressing dramatically. This is caused by progress of the sequencer used for analyses of the DNA sequences. The analyses of the DNA sequences is important for the medical field, agriculture, environmental issues, and so on. Improvement in the speed of the DNA sequencing technique leads to improvement in the speed of the biogenetics. One of the reasons of recent progresses the DNA sequencer is development of shotgun sequencing method.

The shotgun sequencing method is one of the methods which estimate a long DNA sequence from a lot of fragments. A single sequencer cannot read a quite long sequence with a huge number of bases into a computer directly in a short time. Therefore, in the shotgun sequencing method, first, the DNA clones are generated using a vector. Then, the DNA sequence is cut into small fragments using a restriction enzyme or a physical shearing force. The sequencer inputs such small fragments into the computer, in which the original sequence is reconstructed by an assembly algorithm. In the reading phase, the fragments are cut randomly, and they may have the overlaps. Therefore, the original sequence can be estimated by the assembly algorithms by connecting them according to the overlaps. Such an optimal fragment arrangement search becomes an optimization problem [1], which can be formulated as a modified version of the Asymmetric Travelling Salesman Problem (ATSP).

The Traveling Salesperson Problem (TSP) is to find a minimum length tour that visits each city exactly once when a list of cities and their traveling costs are given. Since the TSP and the ATSP belongs to a class of NP-hard, various heuristic algorithms to find near optimal solutions have been proposed for the TSPs and the ATSPs.

In this paper, various heuristic methods are applied to the ATSPs and the DNA fragment assembly problems, to investigate effective algorithm. We evaluate the performance of the Hill Climbing Method, the Simulated Annealing[2], the Tabu Search[3], and the Genetic Algorithms[4], with three types of local search heuristics. Based on the results on the ATSP, those methods are applied to the DNA fragment assembly problem.

2. The DNA fragment assembly problem

The DNA sequencer used in this research generates a lot of small fragments with some overlaps as shown in Fig.1. According to the overlaps of the fragments, they can be connected like a puzzle when reconstructing the original sequence. This is an optimization problem to rearranges such mixed fragments into the right order. We apply optimization methods for the ATSP to this problem, which finds the right order of the cities. The cost in the DNA fragment assembly problem is calculated by comparing the arrangements around the ends of the fragments. It is called the overlap which is calculated by the semi global alignment method [5].

3. Application of Heuristic Methods to Asymmetric Travelling Salesman Problems

The cost function of the TSP is the total length of the tour, which is formulated as follows,

$$f_{P(r)} = \sum_{k=1}^{n} d_{r(\bmod(k)),r(\bmod(k+1))}, \quad (1)$$

where $d_{i,j}$ is the cost when the salesman goes to the city j from the city i, $r(k)$ is the city visited at the k th order, n is the number of cities, respectively. In ATSP, it is not necessarily $d_{i,j} = d_{j,i}$.

In this paper, to investigate effective method for the ATSP, we apply various heuristic methods based on three types of updating on two different definitions of neighboring solution. As the definition of the neighboring solutions, 2-exchange(2-ex) and two types of blockshift operations are introduced.

- 2-exchange (2-ex)

  In the 2-exchange, the neighboring solutions are generated by a simple exchange of two cities. In this paper, we defined it as follows, when (i, j) is selected, the city j will be move to the position after the city i. For example, when $(1, 5)$ is chosen for $(1, 8, 9, 6, 4, 7, 5, 2, 10)$, it is updated to $(1, 5, 9, 6, 4, 7, 6, 2, 10)$.

- BlockShift (BS)

  In the blockshift operation, a selected block is moved to the other position, with keeping the order in the block. In our definition, when (i, j) is selected, the block starting from the city j is move just after the city i. In this paper, we use two types of the blocksizes. The first one(BS1) fixes it at 3, and the second one(BS2) uses the best size at each iteration. For example, when $(1, 5)$ is chosen for $(1, 8, 9, 6, 4, 7, 5, 2, 10)$, it is updated to $(1, 5, 2, 10, 9, 6, 4, 7, 6)$.

As the solution improvement methods applied to the updating method to the neighboring solutions defined above, we introduce the following three algorithms, the hill climbing method(HC), the simulated annealing(SA), and the tabu search(TS).

- Hill Climbing Method (HC)

  When the better neighboring solution is found, the solution is immediately updated to the neighbor. However, this approach has a local minimum problem.

- Simulated Annealing (SA)

  The solution is updated when Eq.(2) is satisfied,

$$rand < \frac{1}{1 + e^{-\frac{\Delta_{i,j}}{T}}}, \quad (2)$$

  where $\Delta_{i,j}$ is the improvement of the objective function by moving to the neighbor (i, j), T is the temperature, *rand* is the uniformly distributed random number between 0 and 1. Whenever iteration increases, temperature T is lowered, and the probability of selecting the better solution is increased.

- Tabu Search (TS)

  Once a move (i, j) is applied to updating the solution, the corresponding move is memorized in a tabu list. The moves in the tabu list are forbidden for a fixed term, called a tabu tenure. The TS updates the solution to the neighbor (i, j), which has the largest $\Delta_{i,j}$ in non-tabu ones.

As a globally searching heuristic method, we introduce the genetic algorithm(GA), which has been applied to DNA fragment assembly problem in Ref. [4].

- Genetic Algorithm (GA)

  First, many solutions are created at random. Two solutions are selected from the set of the current solutions, and crossover operation is applied to create new solutions. Then, the worst ones in all of solutions in the current set are removed, that is called mutation. The crossover operation and the mutation procedures are repeated, and better solutions are found by such mixing of better solutions.

To evaluated performance of three algorithms on ATSPs, we generated benchmark problems whose sizes are 10 to 70, by combining eil76 and st70 in TSPLIB[10]. The results on the ATSPs are shown in Table 1. For each algorithm, the initial solution is generated by greedy algorithm connecting small cost links.

Table 1 shows that the GA exhibits the best only for the 10 city problem and the BS2 updated by the TS is the best for all other larger problems. The blockshift is better updating method to the neighboring solution. The TS is the better the SA for update decision method. From these results, we confirm that the BS2 and the TS is the best for ATSP in these algorithms.

4. Application of Heuristic Methods to the DNA Fragment Assembly Problem

In this research, the DNA fragment assembly problem is solved in a similar way to the ATSPs. A fragment is corresponding to a city, and an overlap to a cost. The cost function of the DNA fragment assembly problem is the total length of the sequence which is formulated as follows,

Table 1: The simulation result of various heuristic methods, based on the two exchange, the block shift, the hill climbing, the simulated annealing, the tabu search, and the genetic algorithm, for ATSPs whose sizes are 10 to 70.

	10	20	30	40	50	60	70
2-ex with HC	161.5	307.9	432.5	503.1	563.2	644.4	738.6
2-ex with SA	147.2	286.3	414.1	475.2	510.7	620.6	721.9
2-ex with TS	143.2	274.2	402.6	460.3	503.8	609.8	700.4
BS1 with HC	165.5	313.3	431.6	489.9	541	634.4	723.5
BS1 with SA	146.5	286.7	415.7	462.4	498.8	617.2	681.7
BS1 with TS	150.6	277.1	387.5	443.9	478.2	593.5	662.2
BS2 with HS	147.1	276.8	401.9	451.3	498.2	603.2	698.7
BS2 with SA	145.1	275.6	400.4	444.3	479.0	581.4	640.5
BS2 with TS	146.8	**271.6**	**384.9**	**439.5**	**448.2**	**555.2**	**617.7**
GA	**140.5**	283.6	451.6	522.5	689.5	765.0	904.0

$$f_{D(r)} = \sum_{k=1}^{n} l_k - \sum_{k=1}^{n-1} overlap_{r(k),r(k+1)}, \quad (3)$$

where $overlap_{i,j}$ is the overlap for connecting the fragment j next to the fragment i, $r(k)$ is the fragment in the k th order, n is the number of fragments, respectively. The overlaps between two fragments are calculated by semi-global alignment method[5]. The target of the DNA fragment assembly problem is to search a shotest sequence by including longer overlaps. The shortest sequence may almost the same as the original sequence.

We have chosen three sequences from the NCBI[11]: a human MHC class II region DNA with fibronectin type II repeats HUMMHCFIB, with accession number X60189, which is 3835 bases long; a human apolipoprotein HUMAPOBF, with accession number M15421, which is 10,089 bases long; and the complete genome of bacteriophage lambda, with accession number J02459, which is 20,014 bases long. By amplifying each genome data and fragmenting them, we generated the DNA fragment assembly problems shown in Table 2. The rate of amplification of each genome data is called a coverage.

We apply ten heuristic methods used in the previous section to these DNA fragment assembly problems, and estimate the original sequence using the order of fragments obtained by the optimization algorithms. To evaluate the correctness of solution, the obtained sequences by each optimization algorithm are compared with the original sequence using the global alignment [6] and a similar rate is calculated, which is defined as follows,

$$SimilarRate \frac{AS_{S\,in,S\,out}}{AS_{S\,in,S\,in}} \times 100(\%), \quad (4)$$

where $AS_{S\,in,S\,out}$ is the alignment score which compares the input sequence with the output sequence, $AS_{S\,in,S\,in}$ is the alignment score which compares the input sequence with itself. The shortest sequence found by minimizing the objectve function in Eq.3 may have higher similarity to the original sequences.

The results of the heuristic algorithm applied to the ATSP are shown in Table 3 by the similar rate. Table 3 shows that the BS2 updated by the TS is the best for all problems. This algorithm was also the best for the ATSP in previous section. From these results, we confirm that the BS2 updated by the TS is the best for the DNA fragment assembly problem in the heuristic algorithms introduced in the paper.

Table 2: The DNA fragment assembly problems used in this paper.

	X60189		M15421		J02459
Coverage	5	6	5	7	7
Number of fragments	60	72	125	175	350

5. Conclusions

In this research, we examined the effectiveness of the heuristics methods of the ATSP for the DNA fragmentation assembly problems. First, we applied various heuristic methods to the benchmark ATSPs. Our results shows that the blockshift operation updated by the TS is better than the GA and the SA for the ATSPs whole size are larger than 20. Next, we applied those heuristics methods to the DNA fragmentation assembly problems. Our results show that the same algorithm the blockshift with the TS exhibits better than other algorithms for all problems. From these results we confirm that the BS2 updated by the TS is effective for the DNA fragmentation assembly problem.

As future works, we will apply the chaotic algorithm [9] to DNA fragment assembly problem. Since the chaotic algorithms have been shown more effective than the TS, it may be even better algorithm for finding optimal se-

Table 3: The simulation result of various heuristic methods, based on the two exchange, the block shift, the hill climbing, the simulated annealing, the tabu search, and the genetic algorithm, for five DNA fragment assembly problems. The results are evaluated by the *SimilarRate*.

	X60189(5)	X60189(6)	M15421(5)	M15421(7)	J02459(7)
2-ex with HC	87.318	86.897	85.517	81.781	72.451
2-ex with SA	89.533	89.039	87.053	85.437	76.224
2-ex with TS	89.402	89.414	87.269	86.443	76.394
BS1 with HC	88.226	87.237	85.699	81.973	74.968
BS1 with SA	90.372	89.131	87.089	85.737	76.249
BS1 with TS	90.024	88.960	87.273	86.895	78.512
BS2 with HS	97.850	97.414	89.068	86.278	80.391
BS2 with SA	97.159	97.550	88.977	86.767	83.682
BS2 with TS	**98.909**	**97.842**	**90.029**	**88.728**	**85.293**
GA	89.125	86.598	80.234	76.910	70.167

quences.

References

[1] G. Minetti, E. Alba, G. Luque, "Seeding strategies and recombination operators for solving the DNA fragment assembly problem," February 1 ,2008

[2] S. Kirkpatrick, C. D. Gelatt Jr. 1, M. P. Vecchi, "Optimization by Simulated Annealing," *Scince*, vol. 220, No. 4598, pp. 671-680, 1983.

[3] E. Taillard, "Robust taboo search for the QAP," *Parallel Computing*, Vol. 17, pp. 443-455, 1991.

[4] J.Grefenstette, R.Gopal, B.Rosmaita, and D.Van Gucht, "Genetic Algorithm for the Traveling Salesman Problem, Proc. of 1st Int. Conf. on Genetic Algorithms and their applications," pp.160-168, 1985.

[5] J.C.Setubal, J.Meidanis, "Introduction to Computational Molecular Biology," *PWS Publishing*, 2001.

[6] N. C.Jones, P. A. Pevzner, "An Introduction to Bioinformatics Algorithms," *The MIT Press*, 2007.

[7] T. Tachibana, M. Adachi, "Solving Asymmetric TSP by Combination of Chaotic Neurodynamics and Block Shift operations," *EICE technical report.*, Nonlinear problems 109(30) pp.63-66, 2009.

[8] T. Matsuura, T. Ikeguchi, "Refractory Effects of Chaotic Neurodynamics for Finding Motifs from DNA Sequences," *IDEAL 2006*, LNCS 4224, pp. 1103-1110, 2006.

[9] M. Hasegawa, T. Ikeguchi, K. Aihara, K. Itoh, "A Novel Chaotic Search for Quadratic Assignment Problems," *European Journal of Operational Research* , vol.139, pp.543—556, 2002

[10] TSPLIB (http://elib.zib.de/pub/mp-testdata/tsp/tsplib/tsplib.html)

[11] NCBI (http://www.ncbi.nlm.nih.gov/)

[12] NITE (http://www.bio.nite.go.jp/ngac/analysis1.html)