The Seventeenth International Symposium on Artificial Life and Robotics 2012 (AROB 17th '12),
B-Con Plaza, Beppu, Oita, Japan, January 19-21, 2012

DNA Sequencing by Max-Min Ant System and Genetic Algorithm

Tao Liu[1] and Michiharu Maeda[2]

Fukuoka Institute of Technology, Fukuoka 811-0295, Japan

[1]mfm10022@bene.fit.ac.jp
[2]maeda@ fit.ac.jp

Abstract: This paper is concerned with DNA sequencing by hybridization. An algorithm that Max-Min Ant System and Genetic Algorithm (MMASGA) is proposed to solve the computational field of sequencing resulting from hybridization experiment. For avoiding the local minimum, MMASGA is based on Max-Min Ant System (MMAS), into which Genetic Algorithm (GA) is added. Before getting into the maximum iteration of MMAS, GA takes place in MMAS. In the numerical evaluation, with the iteration mounting up, the summation of the DNA probe appearance frequency turns more. The accuracy of the solution gotten by MMASGA is higher than MMAS.
Keywords: DNA sequencing, sequencing by hybridization, Max-Min Ant System and Genetic Algorithm.

1 INTRODUCTION

Bioinformatics is the application of computer science and information technology to the field of biology and medicine. DNA sequencing is one of the research objects of the bioinformatics. One of the sequencing methods is sequencing by hybridization (SBH). It is concerned with the computational part of the SBH in this paper. In the case of ideal hybridization experiment, an original DNA sequence can be reconstructed. SBH whose errors are under consideration is regarded as integer program problem and has been formulated. The application of Branch-and-Bound Algorithm for seeking the optimized solution has succeeded, but if there are many errors or the DNA sequence is long, it will take an immense amount of time and the optimized solution cannot be gotten [1].

For seeking the optimized solution Ant Colony Optimization (ACO) has been proposed. ACO is a probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs. Being based on ACO, Max-Min Ant System (MMAS) is proposed and has been applied to the combinatorial optimization problem such as traveling salesman problem (TSP), quadratic assignment problem (QAP), vehicle routing problem (VRP), and internet packet switching.

This paper is concerned with DNA sequencing and Max-Min Ant System and Genetic Algorithm is proposed to solve the computational field of sequencing by hybridization. For avoiding the local solution, the proposed MMASGA is based on MMAS and GA is added into MMAS. Furthermore, by the performance evaluation test, the accuracy of the proposed method is better than the traditional method.

2 Sequencing by Hybridization

2.1 About DNA Sequencing

DNA is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms. The four bases founded in DNA are adenine (A), cytosine (C), guanine (G) and thymine (T). Sequencing by Hybridization is a class of methods for determining the order in which nucleotides occur on a strand of DNA, typically used for looking for small changes relative to a known DNA sequence. The binding of one strand of DNA to its complementary strand in the DNA double-helix is sensitive to even single-base mismatches when the hybrid region is short or if specialized mismatch detection proteins are present. This is exploited in a variety of ways, most notable via DNA chips or microarrays with thousands to billions of synthetic oligonucleotides found in a genome of interest plus many known variations or even all possible single-base variations [3].

In molecular biology, a hybridization probe is a fragment of DNA or RNA of variable length, which is used in DNA or RNA samples to detect the presence of nucleotide sequences (the DNA target) that are complementary to the sequence in the probe. The probe thereby hybridizes to single-stranded nucleic acid whose base sequence allows probe-target base pairing due to complementarity between the probe and target. The labeled probe is first denatured into single stranded DNA (ssDNA) and then hybridized to the target ssDNA immobilized on a membrane [3].

If the gained DNA sequence is ACTCTGG in the hybridization experiment and the probes whose length l is 3 are used, the set of all the probes is {AAA, AAC, AAG… TTT} and the amount of elements is 4^3=64. According to

DNA hybridization, {ACT, CTC, TCT, CTG, TGG} are necessary, as shown in the **Fig.1** [4]**.**

ACT
CTC
TCT
CTG
TGG
ACTCTGG

Fig. 1. Example of sequencing.

2.2 Formulation

If all the probes aren't recognized, the alphabets below *l*– l must overlap in the adjacent probes. And if the wrong probe is not found, several probes must be taken out in the SBH. To deal with all the errors, Blazewics etc. formulated the maximum cardinality found in the SBH below as integer linear programming problem. It is also known as the NP-hard combinatorial optimization problem.

Maximize

$$\sum_{i=1}^{m}\sum_{j=1}^{m} x_{ij} + 1 \qquad (1)$$

Subject to

$$\sum_{i=1}^{m} x_{ik} \leq 1, k = 1, \dots, m \qquad (2)$$

$$\sum_{i=1}^{m} x_{ki} \leq 1, k = 1, \dots, m \qquad (3)$$

$$\sum_{k=1}^{m}\left(\left|\sum_{i=1}^{m} x_{ki} - \sum_{j=1}^{m} x_{jk}\right|\right) = 2 \qquad (4)$$

$$\sum_{s_k \in s'}\left(\sum_{s_i \in s'} x_{ik} . \sum_{s_j \in s'} x_{ki}\right) < |s'| \qquad (5)$$

$\forall$ S' $\subset$ S, S'$\neq$0

$$\sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij}\, x_{ij} \leq n - l \qquad (6)$$

Where m is a cardinality of spectrum, S is a set of probes of cardinality of m, $x_{ij,}$ is a Boolean variable (it is equal 1 if an probe joining vertices i and j is included in the solution; otherwise it is equal to 0), c_{ij} is a cost of a probe joining vertex i with vertex j, and n-*l* is a maximum allowed cost.

Equation (1), to be maximized is equivalent to the number of probes selected. Equation (2) and Equation (3) ensure that each probe included in the solution has at most one immediate successor and one immediate predecessor in the solution path, respectively. Equation (4), the path cannot be the circuit construction. Equation (5), the same probe cannot be used more than twice. Equation (6), a cost connected with the solution is not greater than a given value (equal to n-*l*) [1].

The problem above is one of the kinds that the visited nodes are maximized at the limited distance. For the SBH, distance is the length of the DNA sequence and the nodes are the probes. If the visited nodes get most, the path will be ACT→CTC→TCT→CTG→TGG [4].

3 Ant Colony Optimization and Genetic Algorithm

3.1 Ant Colony Optimization Algorithm

Ant Colony Optimization (ACO) is a population based approach which has been successfully applied to several fields. As the name suggests, ACO has been inspired by the behavior of real ant colonies, in particular, by their foraging behavior. One of its main ideas is the indirect communication among the individuals of a colony of agents, called artificial ants, based on an analogy with trails of a chemical substance, called pheromone, which real ants use for communication. The artificial pheromone trails are a kind of distributed numeric information which is modified by the ants to reflect their experience accumulated while solving a particular problem.

The original idea comes from observing the exploitation of food resources among ants, in which ants' individually limited cognitive abilities have collectively been able to find the shortest path between a food source and the nest. The first ant finds the food source, via any way, then returns to the nest, leaving behind a trail pheromone Ants indiscriminately follow several possible ways, but the strengthening of the runway makes it more attractive as the shortest route. Ants take the shortest route; long portions of other ways lose their trail pheromones. In a series of experiments on a colony of ants with a choice between two unequal length paths leading to a source of food, biologists have observed that ants tended to use the shortest route [5]**.**

The algorithm involves the movement of a colony of ants through the different states of the problem. Thereby, each such ant incrementally constructs a solution to the problem. When an ant completes a solution, or during the construction phase, the ant evaluates the solution and modifies the trail value on the components used in its solution. This pheromone information will direct the search of the future ants.

3.2 Max-Min Ant System

Max-Min Ant System (MMAS) is one kind of the common extensions of ACO. MMAS is that when the pheromones of are updated, the pheromone trails are limited. The added maximum and minimum amounts are $[\tau_{max},\tau_{min}]$. MMAS thought is the most accurate extension of ACO. MMAS differs in three key aspects from ACO [2]**.**

I. To exploit the best solutions found during an iteration or during the run of the algorithm, after each iteration only one single ant adds pheromone. This ant may be the one which found the best solution in the current iteration (*iteration-best* ant) or the one which found the best solution from the beginning of the trial (*global-best* ant).
II. To avoid stagnation of the search the range of possible pheromone trails on each solution component is limited to an interval $[\tau_{max},\tau_{min}]$.
III. Additionally, we deliberately initialize the pheromone trails to τ_{max}, achieving in this way a higher exploration of solutions at the start of the algorithm.

The scheme of MMAS algorithm is shown in the **Fig.2.**

procedure MMAS algorithm:
  Set parameters, initialize pheromone trails,
    upper and lower limit of pheromone trails
  while (termination condition not met) **do**
    Construct Solutions
    Apply Local Search
    Update Trails
    Limit Trails
  end
end

Fig.2. The scheme of MMAS algorithm.

3.3 Genetic Algorithm (GA)

A genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. GAs belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

GAs are adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetic. The basic concept of GAs is designed to simulate processes in natural system necessary for evolution, specifically those that follow the principles first laid down by Charles Darwin of survival of the fittest. As such they represent an intelligent exploitation of a random search within a defined search space to solve a problem [6].

First pioneered by John Holland in the 60s, Genetic Algorithms has been widely studied, experimented and applied in many fields in engineering worlds. Not only does GA provide an alternative method to solving problem, it consistently outperforms other traditional methods in most of the problems link. Many of the real world problems involved finding optimal parameters, which might prove difficult for traditional methods but ideal for GAs. However, because of its outstanding performance in optimization, GAs have been wrongly regarded as a function optimizer. In fact, there are many ways to view genetic algorithms. Perhaps most users come to GAs looking for a problem solver, but this is a restrictive view

4 Improvement of Max-Min Ant System

It is possible for MMAS to lapse into the local best solution. To raise the precision of MMAS, the problem addressed in this paper is concerned with DNA sequencing and Max-Min Ant System and Genetic Algorithm (MMASGA) is proposed to solve the computational field of sequencing by hybridization. The proposed MMASGA is based on MMAS and GA is added into MMAS. Firstly initial solutions are gotten by MMAS. Then initial solutions of MMAS are treated as the initial solutions of GA in order not to lapse into the local solution. And the method of pheromone secretion is also improved. Only the pheromone of the path of the good solution is raised. After that, inheritance, mutation, selection, and crossover take place and repeat until the termination conditions are fulfilled. Before getting into the maximum search iteration of MMAS, GA takes place in MMAS. The algorithm of MMASGA is shown below. (**Fig.3.**)

procedure MMASGA algorithm:
  Set parameters, initialize pheromone trails,
    upper and lower limit of pheromone trails
  while (termination condition not met) **do**
    Construct Solutions
    Apply Local Search
    while (generation not maximum) **do**
      Inheritance
      Selection
      Crossover
      Mutation
    end
    Update Trails
    Limit Trails
  end
end

Fig.3. The scheme of MMASGA algorithm.

The flow of MMASGA is described below:

I. Initialize the parameters: the maximum iteration T_{max}, the iteration $t \leftarrow 0$, the initial scatter list, and the initial pheromone trails.

II. The paths are formed randomly and the searched paths are putted into the scatter list. At the same time, a record of the scatter list is made. The initial solutions are made to be the initial generation population of individuals of GA.

III. Until the maximum generation, inheritance, selection, crossover and mutation take place among the individuals.

IV. The probes are calculated and the maximum value is saved.

V. The pheromone trails that construct the solutions are limited and updated. A new record of the scatter list is made.

VI. Set $t \leftarrow t+1$. If $t = T_{max}$, the program terminates, otherwise the program goes to the step Ⅱ.

5 Experimental Results

The parameters of MMASGA are shown in the following description. The set of probes is {CTC, TCT, ACT, TGG, CTG}, the amount of agents is 50, the maximum iteration is 50, the mutation rate is 0.7, and the crossover rate is 0.6.

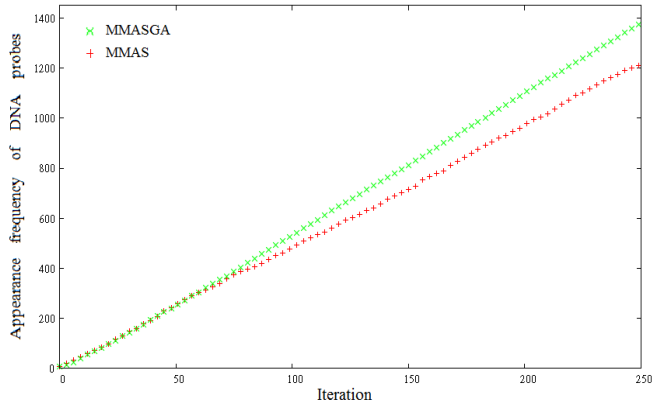

Fig.4. the comparison of performance to MMAS and MMASGA

Table 1. Experimental Results by MMAS and MMASGA

Iteration	MMAS	MMASGA
250	1184	1405
500	2401	3227
1000	4741	6181
10000	47829	79363

The evaluation result is shown above until the iteration counts 250 (**Fig.4.)**. The line of × is the result of calculation by MMASGA. The line of + is the result of calculation by MMAS. With the iteration mounting up, the summation of the DNA probe appearance frequency turns more (**Table 1.**). Through the evaluation of the calculation, the accuracy of the solution gotten by proposed MMASGA is higher than MMAS. When the visited nodes get most, the path turns to be ACT→CTC→TCT→CTG→TGG.

6 Conclusion

In the paper, the Max-Min Ant System and Genetic Algorithm for solving a computational phase of the DNA sequencing by hybridization method has been proposed. The algorithm can handle the DNA sequencing with great accuracy. The presented algorithm has been extensively tasted on real, thus computationally hard, DNA sequences. Parameters n (the length of a reconstructed sequence) and l (the length of hybridizing oligonucleotides) chosen for test purposes, had the values used in real experiments. The tests have shown a good performance of the algorithm in presence. With the application of the global search of GA, the precision of MMAS is raised. Future examinations should allow for a further evaluation of the scope of applications of the sequencing approach with the algorithm proposed in this paper with great accuracy.

REFERENCES

[1] J. Blaewicz, P. Formanowicz, M. Kasprzak, W. T. Markiewicz, and J. Weglarz, "DNA sequencing with positive and negative errors," J. Computational Biology, vol.6, pp. 113-123, 1999

[2] T. Stutzle and H. H. Hoos, "MAX-MIN ant system," Future Generation Computer System, vol.16, no.8, pp. 889-914, 2000.

[3] J. Blazewicz, M. Kasprzak, "Complexity of DNA sequencing by hybridization", Theoretical Computer Science 290 (2003). pp. 1459-1473

[4] Makoto Koshino, Takahiro Otani, Daisuke Taka, and Masatoshi Shirayama, " Applying the Max-Min Ant System to the DNA Sequencing and Its Improvement", The Journal of the institute of Electronics, Information and Communication Engineers, vol. J89-D, no. 5, pp. 911-918, 2006.

[5] M. Dorigo and T. Stutzle, Ant Colony Optimization, MIT Press, 2004.

[6] Schmitt, Lothar M, Theory of Genetic Algorithms, Theoretical Computer Science 259, pp. 1–61, 2001.