Efficient Initial Pool Generation for Weighted Graph Problems Using Parallel Overlap Assembly

Ji Youn Lee[1], Hee-Woong Lim[2], Suk-In Yoo[2], Byoung-Tak Zhang[2], and Tai Hyun Park[1]

[1] School of Chemical Engineering
[2] School of Computer Science and Engineering,Seoul National University, San 56-1 Shilim-Dong, Kwanak-Gu, Seoul 151-744, Korea
elfin94@snu.ac.kr, {hwlim, siyoo, btzhang}@bi.snu.ac.kr, thpark@plaza.snu.ac.kr

Abstract. Most DNA computing algorithms for mathematical problems start with combinatorial generation of an initial pool. Several methods for initial-pool generation have been proposed, including hybridization/ligation and mix/split methods. Here, we implement and compare parallel overlap assembly with the hybridization/ligation method. We applied these methods to the molecular algorithm to solve an instance of the graph problem with weighted edges. Our experimental results show that parallel overlap assembly is a better choice in terms of generation speed and material consumption than the hybridization/ligation method. Simulation of parallel overlap assembly was performed to investigate the potential and the limitation of the method.

1 Introduction

DNA computing has showed its potential by solving several mathematical problems, such as graph and satisfiability problems [1-5]. To solve those problems, precedent initial pool generation is required even though it has been pointed out as a shortcoming in DNA computing. Most molecular algorithms generate initial pools in the first implementation step and then filter the candidate solutions which satisfy the given conditions. Usually, an initial pool is a combinatorial library that contains numerical or indicative information. There are a few initial pool generation methods with their own advantages and disadvantages. One of them is the hybridization/ligation method that link oligonucleotides hybridized through hydrogen bonds by enzymatic reaction. This method was first introduced by Adleman to solve a Hamiltonian path problem [1] and a traveling salesman problem (TSP) [3]. Parallel overlap assembly (POA) was originally introduced by Stemmer to facilitate in vitro mutagenesis [6]. It was successfully applied by Kaplan *et al.* to generate an initial pool consists of binary numbers to solve a maximal clique problem [7]. Other method, such as the mix/split method was introduced by Faulhammer *et al.* to generate combinatorial library of binary numbers [5]. Braich *et al.* applied this method to generate an initial pool for 20-variable 3-SAT problem [4].

In previous work, we implemented a molecular algorithm to solve a 7-city traveling salesman problem [3]. The molecular algorithm for TSP also contained an initial pool

C. Ferretti, G. Mauri and C. Zandron (Eds.): DNA10, LNCS 3384, pp. 215 – 223, 2005.

generation step before filtering and readout steps. The hybridization/ligation method was used to generate an initial pool and we succeeded to solve the problem; however, the pool generation efficiency was very low. Therefore, the hybridization/ligation method cannot guarantee a complete pool as the problem size increases, which limits the solvable problem size. We introduced another initial pool generation method which is based on parallel overlap assembly. This method was compared with the former one by looking at the product size distributions. Additionally, a computerized simulation of parallel overlap assembly was performed to support the experimental results.

2 Initial Pool Generation Methods

2.1 Parallel Overlap Assembly

Kaplan *et al.* suggested the construction of computational DNA libraries based on a DNA shuffling method [2, 6]. They succeeded in constructing a complete library of binary numbers from 0 to 2^4-1 to solve the maximal clique problem for a graph with four vertices. Their library consisted of two parts; one is the position string of fixed length and the other is value string (0 or 1) of various lengths. The DNA strands corresponding to the same position string were overlapped during an annealing step in the assembly process while the remaining parts of the DNA strands were extended by dNTPs incorporation by polymerase (represented by the dotted arrows in Fig. 1).

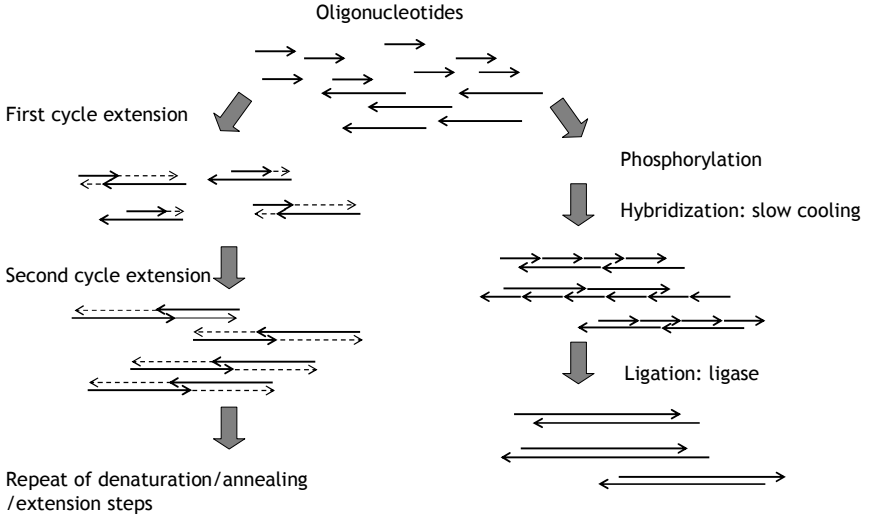

Fig. 1. Schematic diagram of parallel overlap assembly and the hybridization/ligation method for a traveling salesman problem. **Left part**: The thick arrows represent the single-stranded DNA molecules which participate in each cycle of the reaction. The dotted arrows represent the elongated part by dNTPs incorporation. **Right part**: The nicks generated in the hybridization step are linked by ligase via the formation of a phosphodiester bond between the 3' hydroxyl and 5' phosphate of adjacent nucleotides. The arrowhead indicates the 3' hydroxyl end

The mechanism of POA resembles that of polymerase chain reaction (PCR) in that it repeats the denaturation, annealing and extension. However, the characteristics are complete different. PCR is an in vitro DNA amplification method, so the number of target DNA strands doubled every cycle. In POA, the number of DNA strands does not increase as the cycle progresses, while the lengths of the DNA strands increase.

POA can also efficiently be applied to initial pool generation for a weighted graph problem because the encoding scheme relies on hybridization between city sequences which correspond to the positioning string in the maximal clique problem. The schematic diagram of POA for a traveling salesman problem is shown in Fig. 1.

2.2 Hybridization and Ligation Method

Adleman created an initial pool of candidate paths in parallel to solve a 7-node Hamiltonian path problem utilizing the hybridization/ligation method [1]. Possible paths of various lengths were generated by hybridization between half sequences of each node. Ligase connected the nicks between the 3' hydroxyl and 5' phosphate of adjacent nucleotides which are formed after hybridization via a phosphodiester bond and consequently linked the DNA molecules (right part of Fig. 1). This method was also applied to solve a 7-city traveling salesman problem [3].

3 Materials and Methods

3.1 Target Problem

The target problem was a 7-city traveling salesman problem as shown in Fig. 2 (A). DNA strands representing the city and the cost were encoded with 20-mer oligonucleotides. DNA strands representing the road were encoded according to city and cost information with 40-mer oligonucleotides. The last half (10-mer) of the departure city and the first half (10-mer) of the arrival city act as linkers to connect the cities (Fig. 2 (B)). The linker parts hybridize in hybridization and annealing step in each initial pool generation method.

3.2 Parallel Overlap Assembly

Thirty five different DNA oligonucleotides (7 cities, 5 costs, and 23 roads) were mixed and subjected to PCR without primers as templates. The reaction mixture contained 1.25 unit of Pyrobest® DNA polymerase (TaKaRa, Japan) in 10 mM Tris-HCl, pH 9.0, 1 mM $MgCl_2$, 50 mM KCl, and 0.2 mM of each dNTP was dissolved in distilled water. The total reaction volume was 20 µl. PCR was processed for 34 cycles at 95°C for 30 seconds, at 55°C for 30 seconds and at 72°C for 30 seconds. Initial denaturation and prolonged polymerization were executed for 4 minutes each.

3.3 Hybridization and Ligation

The same amount of oligonucleotide mixture as in POA was prepared [3]. The mixture was heated to 95°C and then hybridized by slow cooling to 20°C at 1°C per minute. The reaction mixture was then subjected to a ligation. For a ligation, 5 µl of

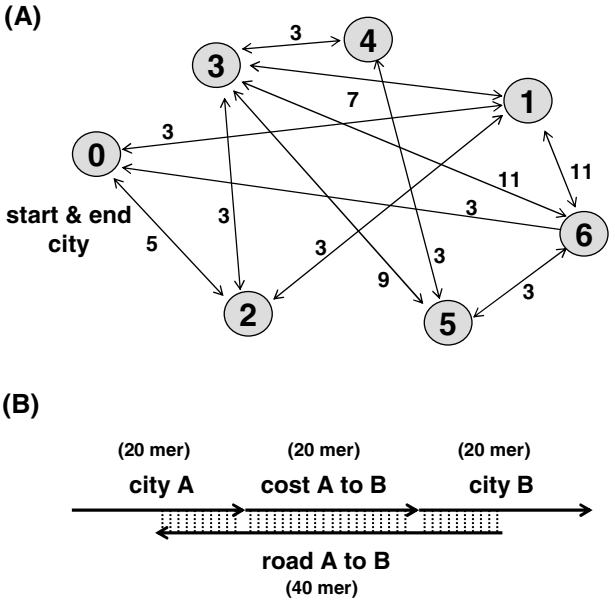

Fig. 2. The seven-city traveling salesman problem (A) and encoding scheme (B). Paths start and end at city 0. The circles denote the cities and the arrows represent the roads. The number on each arrow gives the cost on the given road. The arrowhead of (B) denotes 3' hydroxyl end

the reaction mixtures, 700 units of *T4* DNA ligase (TaKaRa, Japan), ligase buffer (66 mM Tris-HCl, pH 7.6, 6.6 mM $MgCl_2$, 10 mM DTT, 0.1 mM ATP), and an appropriate volume of distilled water was mixed. The total reaction volume was 10 μl. The reaction mixture was incubated at 16°C for 16 hours.

3.4 Gel Electrophoresis and Image Analysis

Agarose gel electrophoresis was performed with 2% Agarose-1000 (Invitrogen, CA, USA) in 0.5X tris-Borate-EDTA buffer and gel was stained with ethidium bromide. As a marker, GeneRuler^{TM} 50 bp DNA ladder (Fermentas, MD, USA) was used. The gel image was obtained with a Gel-Doc and analyzed by Quantity One^{TM} (Bio-Rad, USA).

3.5 Simulation of Parallel Overlap Assembly

We used the algorithm of Maheshri [8] to simulate parallel overlap assembly process for initial pool generation. For simplicity, we only considered match regions to calculate the free energy and did not consider mismatches or dangling ends. The free energy was calculated from the nearest-neighbor model and the parameters given by SantaLucia [9].

4 Results and Discussion

To compare the initial pool generation efficiencies between the hybridization/ligation method and parallel overlap assembly, we performed agarose gel electrophoresis and analyzed the gel image with an image analysis software. The efficiencies can be indirectly compared with the produced amounts of expected length of DNA strands. From the agarose gel, it was possible to determine the length distribution of DNA strands. The length of the candidate paths was 300 bp. The candidate paths contained eight cities and seven costs which were 20-mer respectively; the paths start from city 0, end with city 0, and visit all seven cities. The electrophoresis results are shown in Fig. 3 (A). After the hybridization/ligation reaction, the elongated DNA strands were observed (lane 2 in Fig. 3 (A)) to be located higher than the oligomer mixture (lane 1 in Fig. 3 (A)). The fluorescence intensity was generally increased by the double-stranded DNA formation when compared with oligomer mixture in lane 1. However, most DNA strands are located around 100 bp, which indicates that the generated paths visit only two to four cities. Approximately 13.18 ng of DNA strands were located

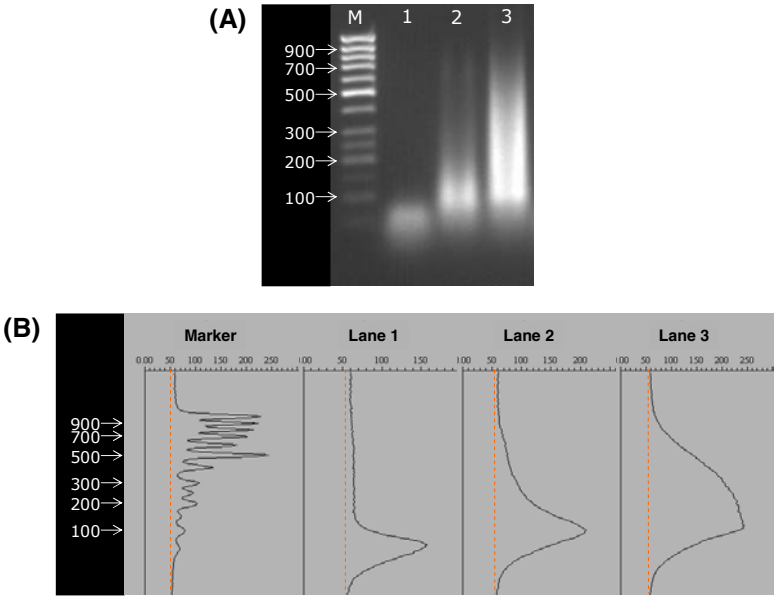

Fig. 3. Comparison of two initial pool generation methods. (A) Experimental results of electrophoresis on 2% agarose gel. M denotes DNA size marker (50 bp ladder). Lane 1 is the oligomer mixture, lane 2 is the product of hybridization/ligation reaction and lane 3 is the product of parallel overlap assembly by *Taq* polymerase. (B) Image analysis results using Quantity One™. Each graph corresponds to each lane of (A). When comparing POA with the hybridization/ligation method, more DNA strands are located around 300 bp which is the expected pool size

around 300 bp, which corresponds to 67.59 fmole molecules or 27.04 nM (quantified by a linear regression of marker DNA molecules). When considering the reactionvolume (10 μl), the generated pool size was 1.63×10^{11}. The complete initial pool size of the 7-city TSP cannot exceed 8^8 (= 1.68×10^7). Therefore, we can conclude that the initial pool generated by hybridization/ligation contained the complete pool.

In the case of parallel overlap assembly, many longer DNA strands were observed compared to the hybridization/ligation reaction (lane 3 in Fig. 3 (A)). Moreover, the product amount that was represented by the area of the peak was increased by the dNTPs incorporation by polymerase. There were approximately 25.82 ng of DNA strands around 300 bp, which corresponds to 132.41 fmoles or 26.48 nM. When considering the reaction volume (20 μl), the generated pool size was 3.19×10^{11}. The initial pool size generated from the same amount of initial oligonucleotides was about two times larger than that of hybridization/ligation. With larger problem, the initial pool size is too small to contain the complete pool; however, POA with more cycle and large experimental scale can include practical pools.

Parallel overlap assembly is a better initial pool generation method for problems that require combinatorial initial pools, such as weighted graph problems. Firstly, POA is more efficient than the hybridization/ligation method in that it maintains the population size, *i.e*. the number of DNA molecules throughout the procedure. Initially, two single-stranded DNA molecules partially hybridize in the annealing step and then they are extended by dNTPs incorporation by polymerase. The elongated DNA molecules are denatured to two single-stranded DNA in the next denaturation step, and they are subjected to the annealing reaction at the next cycle. Therefore, the population size does not change, and we can decide the population size by varying the initial number of oligonucleotides. On the other hand, in the hybridization/ligation method, the population size decreases as reaction progresses. For example, in our target problem, one complete double-stranded DNA strand composed of eight cities can be made by a ligation of eight city strands, seven cost strands, and seven road strands. This means that the population size is decreased by a factor of the number of components composing it in the hybridization/ligation method. As the problem size increase, the required initial pool size increases dramatically. Therefore, in the light of scalability, POA has an advantage over the hybridization/ligation method.

Secondly, POA does not require phosphorylation of oligonucleotides which is prerequisite for the ligation of oligonucleotides. We used 5'-phosphate group modified oligonucleotides for ligation and the oligonucleotide synthesis cost take up most of the expenses. Thirdly, POA demands less time than the hybridization/ligation method. Hybridization required one and half hour while ligation required more than 12 hours; however POA for 34 cycles required only two hours. Therefore, POA is a much more efficient and economic method for initial pool generation. Moreover, initial pool generation by POA requires fewer strands than the hybridization/ligation method to obtain a similar amount of initial pool DNA molecules, because complementary strands are automatically extended by polymerase. For example, in the above target problem, the cost strands are not required because the cost regions can be filled by polymerase extension.

However, POA was not as efficient as we expected. So, we performed a computerized simulation to investigate the capability and the limitation of POA as an initial pool generation method. In the simulation, each cycle consisted of three steps: two single-stranded DNA molecules were randomly selected and collided; annealing event and duplex formation were decided based on the thermodynamic properties; and extendable duplexes were extended according to the pre-determined polymerase fidelity. The selection probability was proportional to the concentration of each DNA strand. Whether single-stranded DNA molecules will hybridize or not was determined by Boltzman-weighted probability.

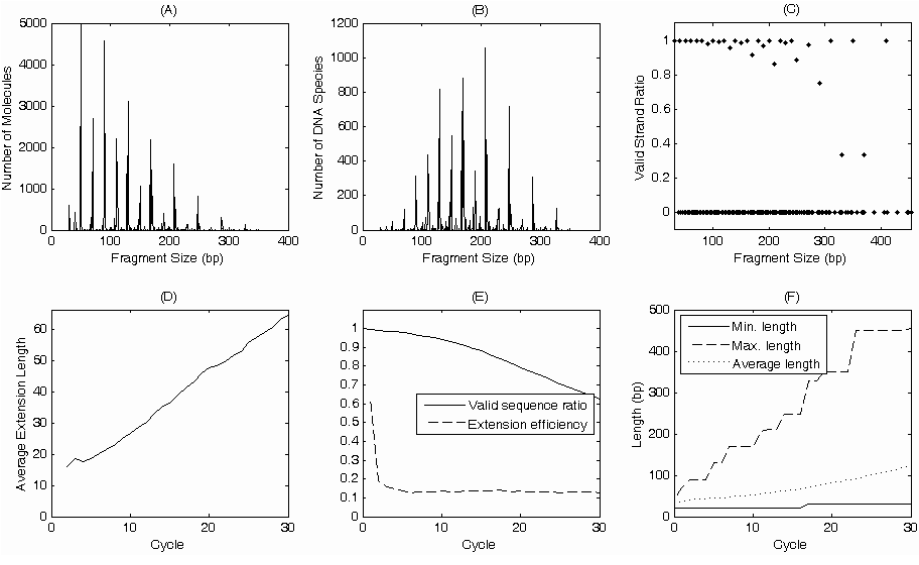

Fig. 4. Simulation results. (A) Length distribution of POA product. (B) Diversity of generated DNA species. (C) Valid strand ratio of final product in each length. (D) Average extension length in each cycle. (E) Valid sequence ratio and extension efficiency in each cycle. (F) Length change during POA: minimum, maximum, and average length in population

After 30 cycles of POA process, various sizes of double-stranded DNA strands were generated (Fig. 4 (A)). The average extension length continuously increased with each cycle (Fig. 4 (D)), which is because extended DNA strands were used as templates for longer strand generation. However, the ratio of DNA strands, which were long enough to contain optimal solution, was very low. This was mainly due to the rapid decrease of the extension efficiency. When we defined the extension efficiency of each cycle as the ratio of annealing event which forms extendable duplex to total annealing event, the extension efficiency dropped dramatically after only a few cycles (Fig 4 (E)). The reason is that non-extendable annealing event increased with cycle. In the early cycles of POA, most annealing between single-stranded DNA molecules formed an extendable duplex. Extendable duplex means that

a 3'-end part of one single-stranded DNA molecule is annealed to its complementary part of the other single-stranded DNA molecule, which can undergo extension by polymerase. However, non-extendable duplex formation rapidly increased with each cycle. Non-extendable duplex means that their 3'-end parts are both dangling, so polymerase cannot incorporate dNTP molecules. The probability of annealing event with a dangling 3'-end increased with each cycle, because the elongated DNA strands had long subsequence complementary to the other strand. They underwent re-annealing rather than initiated an extension reaction. This explained the rapid decrease of the extension efficiency.

Products of POA were diverse both in length and in composition as shown in (Fig 4 (B)). However, the valid strand ratio decreased with each cycle, because the DNA strands extended by mis-hybridization were not eliminated or recovered during the process. Elongated DNA strands can be considered as concatenation of the predefined DNA blocks: left half and right half of city strands, cost strands, and their complementary strands (the region connected by dotted lines in Fig. 2 (B)). Valid strands must be the concatenation of the above DNA blocks. However, invalid strands which are extended after mis-hybridization must contain incomplete subsequences of DNA blocks, and this cannot be recovered during the POA process. Moreover, as DNA strands are extended and getting longer, the possibility of a non-specific annealing increases. These are the reasons why the valid sequence ratio decreased with every cycle, and the valid strand ratio of longer strands were lower than that of shorter strands as shown in Fig. 4 (C). For example, when we investigated the invalid strand of 47 bp after first cycle, we could observe an initial dimer formation between city5→city3 road sequences, which caused incomplete DNA block of 7 bp. We could find the same block in the middle of invalid strands of 287 bp after 30 cycles. This block was found at different positions among those strands. This means that this type of mis-hybridization also happened in a later cycle, because the extension of DNA strands is unidirectional. Like this example, invalid strands generated by mis-hybridization accumulate during POA cycles. Unlike in PCR, which primer strands exist in excess, template strands behave as primers in POA, therefore elaborate 3'-end sequence design is critical for successful POA.

Though POA is a better method for initial pool generation than hybridization/ligation method as mentioned in previous section, the efficiency of POA is not enough to be applied to initial pool generation. Sequence design for initial strands of POA is an important factor and especially the specificity of 3'-end must be considered carefully to prevent an extension after non-specific annealing. In addition, we have to incorporate POA with another supplementary process such as gel electrophoresis and additional amplification of target length by PCR.

Acknowledgements

This research was supported in part by the Ministry of Commerce, Industry and Energy through MEC project, the Ministry of Education & Human Resources Development under the BK21-IT Program and the Ministry of Science and Technology through the NRL Program. The ICT at Seoul National University provided research facilities for this study.

References

[1] L. Adleman. Molecular computation of solutions to combinatorial problems. Science, 266:1021-1024, 1994.
[2] Q. Ouyang, P. D. Kaplan, S. Liu, and A Libchaber. DNA solution of the maximal clique problem. Science, 278:446-449, 1997.
[3] J. Y. Lee, S. –Y. Shin, T. H. Park, and B. –T. Zhang. Temperature gradient-based DNA computing for graph problems with weighted edges. Lect. Notes. Compt. Sci. 2568:73-84, 2003.
[4] R. S. Braich, N. Chelyapov, C. Johnson, P. W. K. Rothemund, and L Adleman. Solution of a 20-variable 3-SAT problem on a DNA computer. Science, 296:499-502, 2002.
[5] D. Faulhammer, A. R. Cukras, R. J. Lipton, and Laura F. Landweber. Molecular computation: RNA solutions to chess problems. Proc. Natl. Acad. Sci. USA. 98: 1385-1389, 2000.
[6] W. Stemmer. DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. Proc. Natl. Acad. Sci. USA. 91: 10747-10751, 1994.
[7] P. D. Kaplan, Q. O. Ouyang, D. S. Thaler, and A Libchaber. Parallel overlap assembly for the construction of computational DNA libraries. J. Theor. Biol., 188: 333-341, 1997.
[8] N. Maheshri, and D. V. Schaffer. Computational and experimental analysis of DNA shuffling. Proc. Natl. Acad. Sci. USA. 100: 3071-3076, 2003.
[9] J. SantaLucia, Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc. Natl. Acad. Sci. USA. 95: 1460-1465, 1998