Finding all the Shortest Paths of a Graph by a DNA-Based Computing Algorithm

Zuwairie Ibrahim, Yusei Tsuboi, Osamu Ono, and Marzuki Khalid

Institute of Applied DNA Computing, Meiji University, Kanagawa, Japan
e-mail: zuwairie@ieee.org, tsuboi@isc.meiji.ac.jp, ono@isc.meiji.ac.jp, marzuki@utmkl.utm.my

***Abstract*—A DNA-based computing algorithm for finding all the shortest paths of a graph is presented, which returns the minimum weight for every pair of vertices in a graph. During the computation, each shortest path is searched by extracting the shortest DNA strand after gel electrophoresis, which is also crucial in order to "print" the result of the computation.**

***Index Terms*— DNA computing, molecular manipulation, all pairs shortest path, graph problem.**

I. Introduction

DNA computing is a new research of interest where biological tools are employed as a new medium for computation. As of 1994, Leonard M. Adleman [1-2] launched a novel evolutionary approach to solve the Hamiltonian Path Problem (HPP) with seven vertices by DNA molecules. While in conventional silicon-based computer, information is stored as binary numbers in silicon-based memory, he encoded the information of the vertices by a randomly DNA sequences. During the computation, gigantic memory capacity and massively parallelism inherent in DNA computing is exploited to make a brute force search on a big problem space in constant or polynomial-time. The output of the computation, also in the form of DNA molecules can be read and printed by electrophoretical fluorescent method.

In this paper, the operation to find the entire shortest paths of a graph, or the so-called all pair shortest path problem, APSP based on DNA computing algorithm is proposed. This is a well-studied graph theoretic problem in transportation and communication networks. Given a graph $G = \langle V,E \rangle$ where V and E are the vertices and edges of a graph respectively, and a positive length function $W : E \rightarrow R$, the APSP problem is to find, for each pair of vertices, υ_i, $\upsilon_j \in V$ the length of the shortest path from υ_i to υ_j [3]. A weighted directed graph is used to model the problem as depicted in Fig. 1. Similarly, Table 1 lists the shortest path for every pair of vertices in matrix representation. The matrix-based representation as in Table 1 can be used to explain the APSP computation by DNA-based computing approach.

As discussed by many researchers, this problem could be solved by applying Dijkstra's algorithm [4] for single source shortest path (SSSP) computation for every vertex as a fixed source vertex. For instance, Dey and Srimani [5] reported a fast parallel algorithm for APSP problem and its VLSI implementation. On the other hand, Pramanick [6] proposed a distributed computing solution based on distributed Dijkstra's algorithm and distributed Floyd's algorithm.

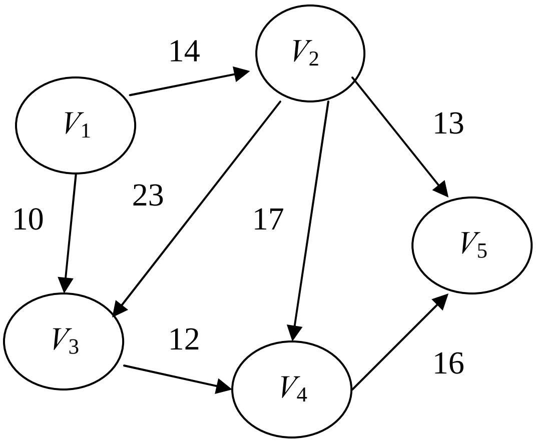


Fig. 1 A weighted directed graph $G = (V, E)$

Table 1. All pairs shortest path representation by a matrix

	V_1	V_2	V_3	V_4	V_5
V_1	-	14	10	22	27
V_2	-	-	23	17	13
V_3	-	-	-	12	28
V_4	-	-	-	-	16
V_5	-	-	-	-	-

Although there has been a lot of discussion regarding the fast and efficient algorithm for APSP problem and its implementation on silicon-based medium for computation, the authors found that there is no, at this moment, research reported on the APSP algorithm based on DNA as medium for computation. From molecular computation perspective, the DNA-based computing approaches reported so far emphasize on the single pair shortest path problem. Hence, this could be the first effort to do such computation by DNA molecules. In this paper, every edge in the graph is modelled by oligonucleotides and the shortest path of every pair could be extracted by means of gel electrophoresis. After the annealing and hybridization operation, the respective path could be amplified by employing PCR operation. Thus for each test tube containing DNA solution, only a certain type of DNA strands will occupy the major part of the solution. The computation returns the minimum weight for every path and the results are "printed" by gel electrophoresis.

II. DNA BIOCHEMICAL OPERATIONS

DNA Synthesis: It is possible to get a solution of DNA molecules with a desired sequence by DNA synthesis. At present, this can be done by ordering the solution from available commercial company.

Denaturation: Double stranded DNA molecules can be separated without breaking the single strands by applying heat to the solution. The double stranded molecules will come apart because the hydrogen bonds between complementary nucleotides are much weaker than the covalent bond between the adjacent nucleotides in the same strands.

Annealing or Hybridization: As oppose to denaturation, by cooling down a DNA, single Watson-Crick complementary DNA sequences will combines to form double stranded molecules. Also, based on this behavior, concatenation of two identical strands would be happened with the presence of a connective strand as depicted in Fig. 2.

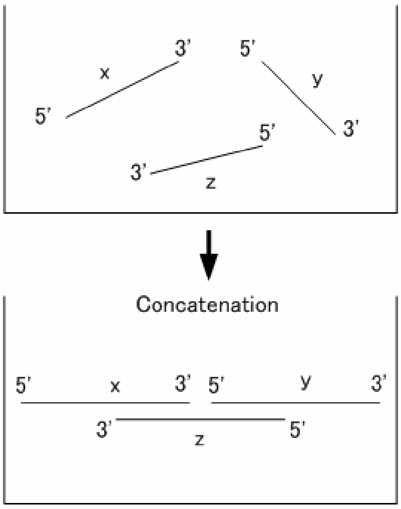


Fig. 2. Concatenation of strand 5' – x – 3' and 5' – y – 3' with a presence of strand 3' – z – 5'

Polymerase Chain Reaction: PCR is an incredible sensitive copying machine for DNA. Given a site-specific single molecule DNA, a million or even billion of similar molecules can be created by PCR process. PCR needs a number of sub-sequence strands called 'primers', which is usually about 20 base long to signal a specific start and end site at a template for replication.

Ligation: Ligation often invoked after the single DNA strands are annealed and concatenated to each other. Many single-strand fragments will be connected in series and ligase is used as 'glue' to seal the covalent bonds between the adjacent fragments.

Gel Electrophoresis: DNA strands in a solution can be separated in term of its length by means of gel electrophoresis. In fact, the molecules are separated according to their weight, which is almost proportional to their length [7]. This technique is based on the fact that DNA molecules are negatively charged [8]. Hence, by putting them in an electric field, they will move towards the positive electrode at different speed. The longer molecules will remain behind the shorter ones. Depending on gel porosity, the precision of gel electrophoresis can be varied, even molecules, which differ by one nucleotide, can still be distinguished between each other. An example of gel electrophoresis output [9-10] is well depicted in Fig. 3. This technique can be used to "print" the results of DNA computation as well.

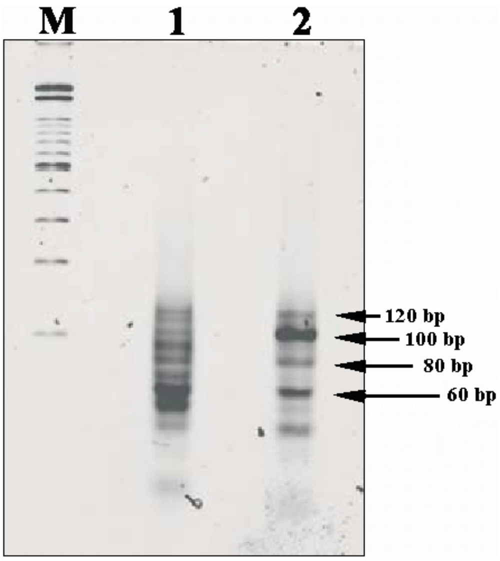


Fig. 3. Gel electrophoresis output where lane M is DNA size marker. Lane 1 and 2 are used for the tested DNA molecules

III. COMPUTING WITH DNA

The overall computation proposed in this paper consists of five steps altogether. Step 1 through step 3 is important for generating the initial pool of random route formation whereas step 4 and step 5 are designed for the computation.

Step 1: Let V be the total number of nodes in the graph and E the total number of paths in the graph. The DNA strands correspond to all nodes and its complements are created. Let O_i (i = 1,....., V) and $\sim O\_i$ (i = 1,....., V) be the fixed length, β random sequences correspond to all nodes in the graph and its complements respectively. For instance, if β = 4, all the oligonucleotides for nodes are placed in Table 2. At the end of this step, oligonucleotides $\sim O\_i$ are synthesized.

Table 2. Fixed length node sequences

Node	DNA Code	Complement
V_1	TATT	ATAA
V_2	GCGG	CGCC
V_3	GTCG	CAGC
V_4	AGAA	TCTT
V_5	TCCG	AGGC

Step 2: In the same way, let O_d (d = 1,....., E) and $\sim O\_d$ (d = 1,....., E) be the variable length random sequences correspond to all the distances and its complements respectively. Let i be the start node of a path, ω the path's distance, α the direct proportional factor, and j the end node of that path. Oligonucleotides representing every path between two nodes in the graph are synthesized as follows:

synthesize oligonucleotides
Rule 1:

$$\textbf{HR}\ O\_i + \{\textbf{ALL}\ O\_d \text{ at length } (\omega\ \alpha - \frac{3\beta}{2})\} + \textbf{ALL}\ O\_j$$

Rule 2:

HR *O_i* + {**ALL** *O_d* at length $(\omega\ \alpha - \beta)$}+ **HL** *O_j*

where '+' is a join, **ALL**, **HR**, and **HL** represent all, half right, and half left of the node sequences respectively. At the end of this step, oligonucleotides ~*O_d* are synthesized.

As such, for the case $\beta = 4$, $\alpha = 2$, if *O_i* = **TATT** then ~*O_i* = **ATAA**, if *O_j* = **GCGG** then ~*O_j* = **CGCC**. If *O_d* at length $(\omega\ \alpha - \frac{3\beta}{2})$ is **AC**, then ~*O_d* = **TG**. If *O_d* at length $(\omega\ \alpha - \beta)$ is **ACGA**, then ~*O_d* = **TGCT**. Hence, according to rule 1, the synthesized oligonucleotide will be **TT+AC+GCGG** = **TTACGCGG**. By the same manner, according to rule 2, the synthesized oligonucleotide will be **TT+ACGA+GC** = **TTACGAGC**.

At the end of this step, 14 identical oligonucleotides for edges will be created by a combination of the nodes sequences and distance sequences. Note that only the number of bases of distance sequences is shown and all these synthesized oligonucleotides are placed in Table 3. These two rules however, decrease the generality of the approach because the weight of edges less than 3 is not allowed, in this case.

Table 3. Oligonucleotides for edges

Node	Oligonucleotides
$V_1 - V_3$	TT[14]GTCG TT[16]GT
$V_3 - V_4$	CG[18]AGAA CG[20]AG
$V_2 - V_5$	GG[20]TCCG GG[22]TC
$V_1 - V_2$	TT[22]GCGG TT[24]GC
$V_4 - V_5$	AA[26]TCCG AA[28]TC
$V_2 - V_4$	GG[28]AGAA GG[30]AG
$V_2 - V_3$	GG[40]GTCG GG[42]GT

Step 3: For the hybridization bio-step, all the oligonucleotides followed by the complement of weight oligonucleotides, ~*O_i* are inserted into a test tube. At this time, hybridization is allowed to occur but according to the content of the solution, only the edge oligonucleotides and the complement of weight oligonucleotides will take part during the hybridization. After the hybridization, for the concatenation bio-step, the complement oligonucleotides of node sequence, ~*O_d* are poured into the test tube. A complement oligonucleotide will concatenate two particular hybridized oligonucleotides as shown in Fig. 4 and as a result, a numerous number of random routes through the graph are formed.

Then DNA ligase reaction is performed. After that, in order to make sure that no sticky end strands exist in the solution, polymerase enzyme is used for the extension where the sticky end strands are altered to become blunt end strands.

Step 4: Produce another four solutions from the initial solution, N_0. Each solution represents exactly one row in Table 1.

Step 5: For each solution (or each row), Select a particular node as a starting point and find the shortest path from the starting node to other nodes. The overall procedures are summarized by a flow chart given in Fig. 5.

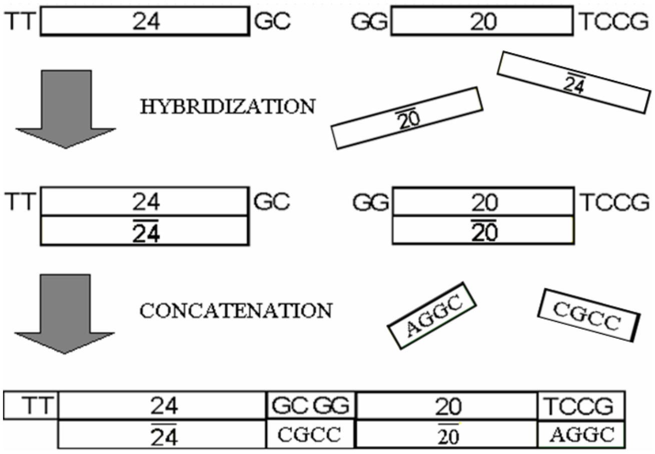


Fig. 4. As an example, hybridization-concatenation bio-steps to the generation of DNA molecules representing the path V_1 - V_2 - V_5

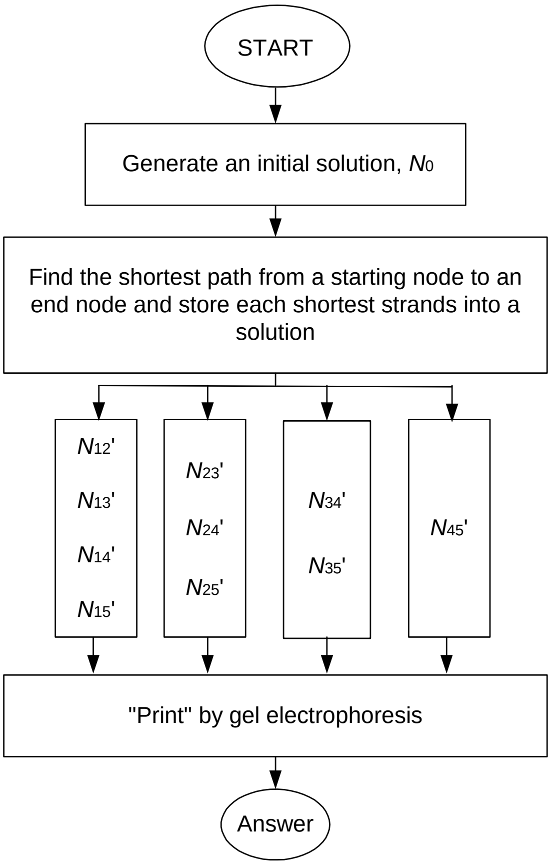

START
Generate an initial solution, N_0
Find the shortest path from a starting node to an end node and store each shortest strands into a solution
N_{12}'
N_{13}'
N_{14}'
N_{15}'
N_{23}'
N_{24}'
N_{25}'
N_{34}'
N_{35}'
N_{45}'
"Print" by gel electrophoresis
Answer

Fig. 5. Overall procedures

IV. DISCUSSIONS

The fourth step could be done by merely divides the solution N_0 equally into five solutions and marks the solutions as N_1, N_2, N_3, and N_4. In this work, it is assumed that all the strands are spread to all areas in the solution.

It is possible to carry out the fifth step by the subsequent procedure. Since the dividing operation is a volume decreasing operation, it is important, therefore, to increase the volume of solutions to a sufficient amount. The only operation, which is able to increase the volume of the solutions, is the polymerase chain reaction (PCR) operation. Besides, this operation selects the routes that begin with a particular node for each test tube obtained in the previous step. Four types of primers will be employed where one primer is assigned to one particular test tube as shown in Fig. 6.

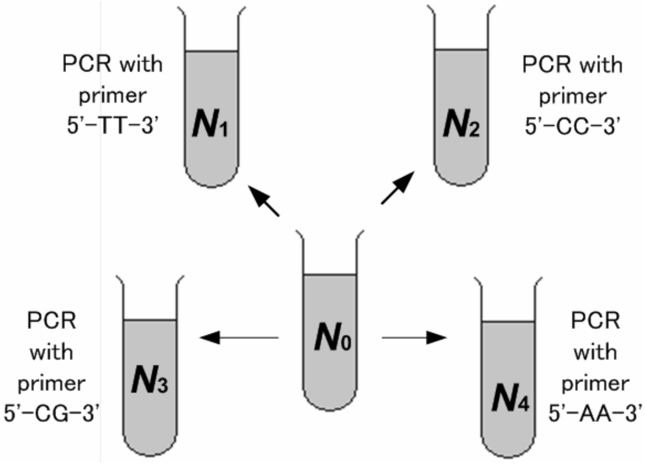


Fig. 6. The initial solution N_0 is divided equally into 5 solutions. PCR process is executed to each solution with 5 different primers representing 5 starting nodes in order to increase the volume of the solutions to a sufficient amount

Next, closely similar to the previous operation, by assuming that all the strands are spread uniformly in the solution, each solution obtained from previous operation, N_1, N_2, and N_3 is divided equally into another 4, 3, and 2 test tubes respectively. As a result, in this case, there should be 10 test tubes altogether. Next, DNA strands with a particular end node are amplified and consequently, the volume of solutions is increase as well. PCR process is employed for each test tube by primers encoding the starting node, O_i, and 3' half complement of the end nodes. This operation is well depicted in Fig. 7.

For better understanding, take a look closely at the molecules of solution N_{12} only. After two PCR operations are accomplished, there should be a numerous number of strands representing the beginning node, V_1 and the end node, V_2. Some of them are shown in Fig. 8.

Next, gel electrophoresis technique will separates these strands in term of length and the shortest length is chosen in order to obtain the shortest path. As a result, the strands $V_1 \rightarrow V_2$ will be chosen. The output solution is called N_{12}'. The same procedure will be executed to the remaining test tube. The expected output "printed" by gel electrophoresis is depicted in Fig. 9.

V. CONCLUSIONS

This paper carried out a DNA-based computing algorithm for all pairs shortest path (APSP) problem of a graph or network. The series of molecular operations are applied to find all the shortest paths exist between two nodes in a graph. During the computation, only the common DNA computing operations such as synthesizing, annealing, hybridization, ligation, PCR, and gel electrophoresis are employed. The contents of this paper also imply the procedures of the laboratory experiment for computation. Since this research is the first attempt to solve APSP problem based on DNA computing, there is still a lot of aspects should be considered and to be explored before the optimal procedure is produced for implementation, in future.

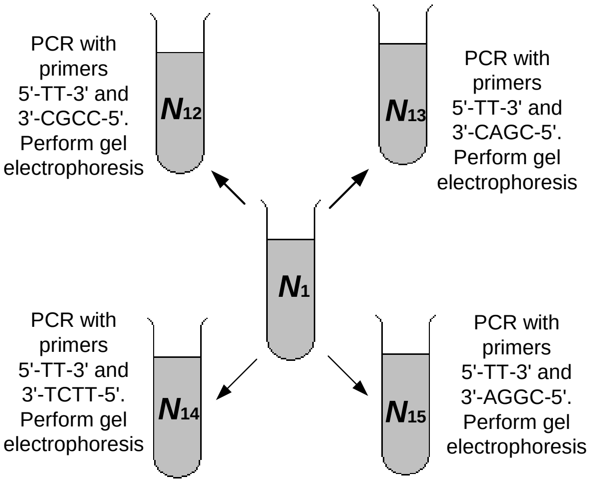
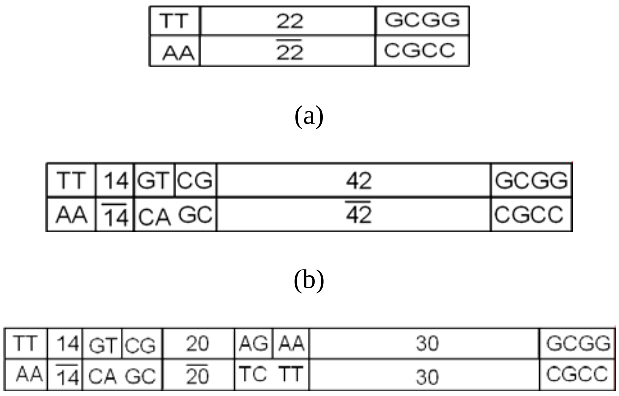

PCR with primers 5'-TT-3' and 3'-CGCC-5'. Perform gel electrophoresis
N12
N13
PCR with primers 5'-TT-3' and 3'-CAGC-5'. Perform gel electrophoresis
N1
PCR with primers 5'-TT-3' and 3'-TCTT-5'. Perform gel electrophoresis
N14
N15
PCR with primers 5'-TT-3' and 3'-AGGC-5'. Perform gel electrophoresis

Fig. 7. For instance, N_1 is divided into another four test tubes namely N_{12}, N_{13}, N_{14}, and N_{15}. PCR process is executed again to each solution with primers representing the 3' half of start node and complement of end nodes. Next, for each test tube, the gel electrophoresis is applied for separating the strands by length



TT	22	GCGG
AA	$\overline{22}$	CGCC

(a)

TT	14	GT	CG	42	GCGG
AA	$\overline{14}$	CA	GC	$\overline{42}$	CGCC

(b)

TT	14	GT	CG	20	AG	AA	30	GCGG
AA	$\overline{14}$	CA	GC	$\overline{20}$	TC	TT	30	CGCC

(c)

Fig. 8. (a) Strand $V_1 \rightarrow V_2$ (b) Strand $V_1 \rightarrow V_3 \rightarrow V_2$, and (c) Strand $V_1 \rightarrow V_3 \rightarrow V_4 \rightarrow V_2$

VI. References

[1] L.Adleman, “Molecular computation of solutions to combinatorial problems,” Science, Vol. 266, pp. 1021-1024, 1994.

[2] L.Adleman, “Computing with DNA,” Scientific American, pp. 34-41, 1998.

[3] N.Alon, Z.Galil, and O.Margalit, “On the exponential of the all pairs shortest path problem,” Journal of computer and system sciences, Vol. 54, No. 2, pp. 255-262, 1997.

[4] E.W.Dijkstra, “A note in connexion with graphs,” Numerische mathematik, Vol. 1, pp. 269-271, 1959.

[5] S.Dey, and P.K.Srimani, “Fast parallel algorithm for all-pairs shortest problem and its VLSI implementation,” IEE Proceeding – Computers and Digital Techniques, Vol. 136, Issue 6, pp. 85-89, 1993.

[6] I.Pramanick, “Distributed computing solutions to the all-pairs shortest path problem,” Second International Symposium on High Performance Distributed Computing, pp. 196-203, 1993.

[7] C.S.Calude, and G.Paun, “Computing with cells and atoms: An introduction to quantum, DNA, and membrane computing,” Taylor & Francis Inc., New York, 2001.

[8] G. Paun, G. Rozenberg, and A. Salooma, “DNA computing: New computing paradigm,” Springer-Verlag, Heidelberg, 1998.

[9] M.Yamamoto, A.Kameda, N.Matsuura, T.Shiba, Y.Kawazoa, and A.Ahochi, “A separation method for DNA computing based on concentration control,” New generation computing, Vol. 20, No. 3, pp. 251-262, 2002.

[10] Y.Yamamoto, A.Kameda, N.Matsuura, T.Shiba, Y.Kawazoe, and A.Ahochi, “Local search by concentration-controlled DNA computing,” International journal of computational intelligence and applications, Vol. 2, No. 4, pp. 447-455, 2002.

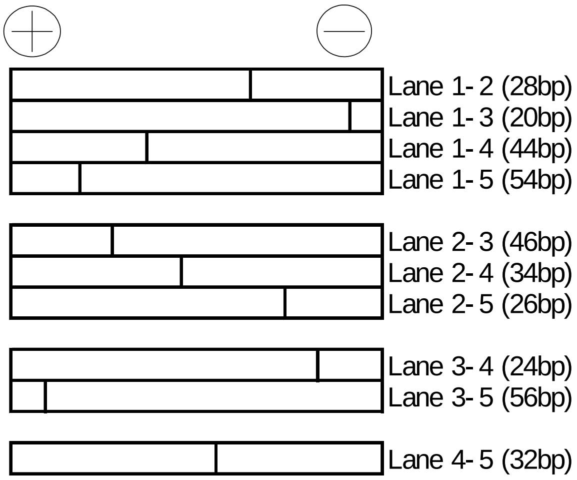


Fig. 9. Expected output ‘printed’ by gel electrophoresis