International Journal of Computer Science & Applications
Vol. I, No. 1, pp. 46 - 60

Direct-Proportional Length-Based DNA Computing for Shortest Path Problem

Zuwairie Ibrahim, Yusei Tsuboi, Osamu Ono

Institute of Applied DNA Computing
Meiji University
1-1-1 Higashi-mita, Tama-ku
Kawasaki-shi, Kanagawa-ken, JAPAN 214-8571
E-mail: (zuwairie, tsuboi, ono)@isc.meiji.ac.jp

Marzuki Khalid

Center for Artificial Intelligence and Robotics (CAIRO)
Universiti Teknologi Malaysia
Jalan Semarak, 54100 Kuala Lumpur, Malaysia
E-mail: marzuki@utmkl.utm.my

Abstract:

Deoxyribonucleic Acid or DNA computing has emerged as an interdisciplinary field that draws together chemistry, molecular biology, computer science, and mathematics. From the DNA computing point of view, it has been proven that it is possible to solve weighted graph problems by exploiting some characteristics of DNA such as length, concentration, and melting temperature. In this paper, we present an alternative direct-proportional length-based DNA computing approach whereby the cost of each path is encoded by the length of the oligonucleotides in a proportional way. The advantage is such that, after the hybridization and ligation reactions, gel electrophoresis can be performed to separate the respective DNA duplex according to their length which directly decodes the results. In addition to this advantage, the reliability of the proposed approach can be enhanced as only the general and well-known bio-molecular laboratory operations are employed during the computation.

Keywords: Length-based DNA computing, shortest path problem, molecular manipulation

1. INTRODUCTION

In 1965, Moore (1965) observed an exponential growth in the number of transistors per integrated circuit against time. Currently, this is the definition of Moore's Law,

meaning that more and more transistors can be crammed into a single chip until the silicon itself reaches its finite atomic scale limitation. Since the traditional silicon-based computer is restricted by its fundamental physical limitation, researchers have been searching for alternative medium for computation and DNA would turn out to be the answer.

DNA molecules are composed of single or double DNA fragments or often called oligonucleotides or strands or oligo in short. These nucleotides form the basis of DNA. A single stranded (ss) fragment has a phospho-sugar backbone and four kinds of bases denoted by the symbols A, T, G, and C for the bases adenine, thymine, guanine, and cytosine respectively. These four nucleic acids, which can occur in any order, combined in Watson-Crick complementary pairs to form the famous double stranded (ds) helix (duplex) of DNA. Due to the hybridization reaction, A is complementary with T and C is complementary with G. As an example, any oligo, such as 5'-ACCTG-3' has a complementary sequence, 3'-TGGAC-5'. Digits 5' and 3' denote orientation of oligo.

A new computing paradigm based on DNA molecules appeared in 1994 when Leonard M. Adleman (Adleman, 1994) launched a novel *in vitro* approach to solve the so-called Hamiltonian Path Problem (HPP) with seven vertices by DNA molecules. The goal of the HPP is to determine whether a path exists that commences at the 'start city' and finishes at the 'end city'; and passes through each of the remaining cities exactly once. While in conventional silicon-based computers, information is stored as binary numbers in silicon-based memories; in this approach, he encoded the information of the vertices by random DNA sequences. The computation is performed in bio-molecular reaction fashion involving procedures such as hybridization, denaturation, ligation, Polymerase Chain Reaction (PCR), etc. The output of the computation, also in the form of DNA molecules can be read and printed by a process called electrophoretical fluorescence.

Four years later, a constant proportional length-based DNA computing technique specifically for Traveling Salesman Problem (TSP) or shortest path HPP is proposed by Narayanan and Zorbalas (Narayanan and Zorbalas, 1998). A constant increase of DNA strands is encoded according to the actual length of the distance. A drawback of this method is that, there is a possibility of an occurrence of concatenated DNA strands of two distances which could be longer than the DNA strand of the longest distance that has been encoded. This may lead to errors in computing the shortest path (Lee *et al.*, 2003). This scheme, however, has not been realized by any laboratory experiment.

Yamamoto (Yamamoto *et al.*, 2002a) presented concentration-controlled DNA computing for accomplishing a local search for the shortest path problem. Although DNA computing with concentration control method enables local search among all the candidate solutions, it cannot guarantee that the most intensive band is the DNA representing the shortest path in the given graph. In addition, it is technically difficult to extract a single optimal solution from the most intensive band (Lee *et al.*, 2003).

Lee (Lee *et al.*, 2002) proposed a DNA computing technique based on temperature gradient for solving the TSP problem. They introduced a melting temperature (T_m) where DNA strands of correct solutions will be denatured and amplified by the Polymerase Chain Reaction (PCR) process. As the denaturation temperature increases, other DNA strands are also subsequently amplified, however, the amount of correct solutions will also be exponentially increased which does affect the final solution.

Due to drawbacks in implementation, the constant proportional length-based DNA computing has not yet been implemented in any laboratory experiment. Thus, with

the aim to solve the limitation of the constant-proportional length-based approach by improving the previously proposed encoding style in (Narayanan and Zorbalas, 1998), a direct proportional length-based DNA computing approach is proposed in this paper. The shortest path problem has been selected for consideration of using the proposed technique. In this approach, the cost of an edge is encoded as a direct-proportional length oligo. As a result, during the computation, the important information is the length of the DNA duplex. Since this will result in numerous numbers of combinations, by using the standard bio-molecular laboratory operations, it is possible to extract the optimal combination which represents a solution to the problem.

2. CONSTANT PROPORTIONAL LENGTH-BASED DNA COMPUTING

We first briefly discuss the constant proportional length-based DNA computing technique. Consider a directed graph as shown in Figure 1. The example graph consists of four nodes, four edges, and costs. If the smallest numerical length is encoded by an oligo of length two with the constant proportional increase of two bases, its numerical length-DNA pairs can be listed as in Table 1. If one wants to compute the shortest path from node *a* to node *c*, two kinds of path with different numerical lengths existed as shown in Figure 2.

If the constant proportional length-based approach is further examined, it can be seen that it is possible for some cases, the numerical length of a longer path is encoded by a shorter DNA length and vice versa. Thus, Figure 2 clearly shows that the DNA length for numerical length 19 is larger than the DNA length for a numerical length of 23. The obvious answer is to make the DNA length of a path directly proportional to the distance of that path. This idea leads to a direct proportional length-based DNA computing approach for shortest path problem.

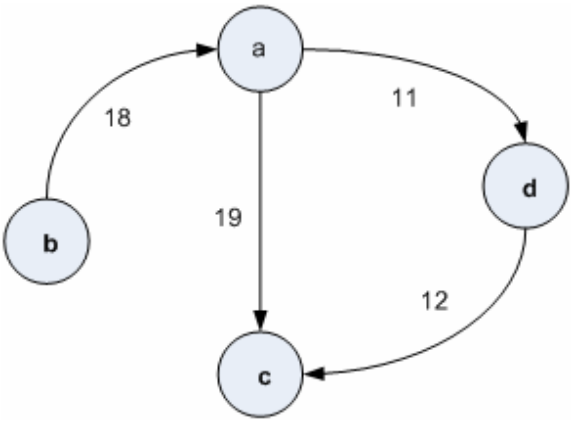

Figure 1. Example of a directed graph of constant proportional length-based DNA computing

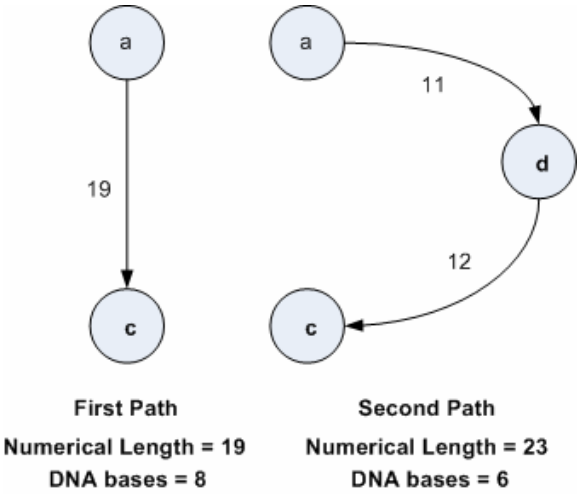

Figure 2. Example of two different paths for the path *a-c*

Table 1. Numerical – DNA length pairs for example as shown in Fig. 1

Numerical length	DNA length
11	2-mer
12	4-mer
18	6-mer
19	8-mer

3. THE SHORTEST PATH PROBLEM

The input to this problem is a weighted (cost), directed graph $G = (V, E, \omega)$, a start node u and an end node v. The output of the shortest path problem is a (u, υ) path with the smallest cost. In the case given in Figure 3, if u is V_1 and υ is V_5, the cost for the shortest path will be given as 27 and the optimal path is clearly shown as $V_1 - V_2 - V_5$.

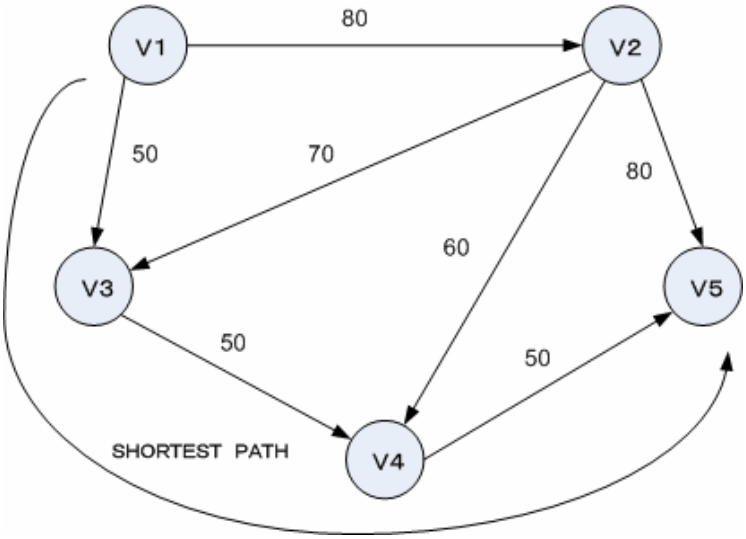

Figure 3. Example showing a weighted directed graph $G = (V, E)$ with the shortest path shown as V_1 - V_3 - V_4 - V_5

4. DNA BIO-MOLECULAR OPERATIONS

There are a number of feasible DNA bio-molecular operations in DNA computing. Such operations are DNA synthesis, polymerase chain reaction (PCR), ligation, and gel electrophoresis. In this section, these essential tools will be described in more detail.

4.1 DNA Synthesis

At present, it is possible to get a test tube containing approximately 10^{18} DNA molecules with desired sequences. Some commercial DNA synthesis companies are available which provide such sequences at a reasonable price. According to Adleman (Adleman, 1998), the cost for DNA sequences of length 20 is just about $25.

4.2 Denaturation

Double stranded (ds) DNA molecules can be separated without breaking the single strands by applying heat to the solution at about 95°C. The double stranded (ds) DNA 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.

4.3 Annealing, Concatenation, and Hybridization

Contrary to denaturation, by cooling down the DNA solution from approximately 95°C to 55°C, single stranded (ss) oligos will combine to form double stranded (ds) duplex; this process is called annealing. Also, based on this behavior, concatenation

of two identical strands takes place with the presence of a connective strand as depicted in Figure 4.

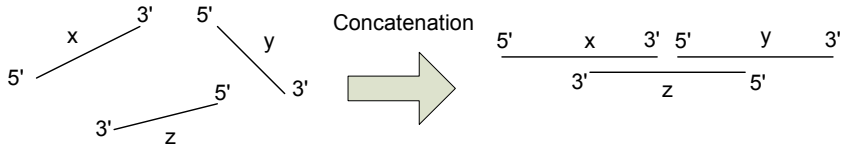

Figure 4. Concatenation of strand 5' – x – 3' and 5' – y – 3' with a presence of connective strand 3' – z – 5'

4.4 Polymerase Chain Reaction (PCR)

PCR is an incredibly sensitive copying machine for DNA. It also can be used for DNA detection. Given a site-specific single molecule DNA, a million or even billion of similar molecules can be created by the PCR process. In n steps, this operation produces 2^n copies of the same molecules. PCR needs a number of sub-sequence strands called 'primers', which is usually about 20 base long to signal a specific start and end sites of a template for replication. PCR normally runs for 20-30 cycles of 3 steps as shown in Figure 5. Each cycle consists of three phases: separating base pair strands of DNA at about 95°C, annealing at 55°C, and extension at 75°C (Fitch, 2002). PCR takes about 3 hours normally to complete the cycles.

4.5 Ligation

Ligation is often invoked after the single DNA strands are annealed and concatenated to each other. Many single strand (ss) fragments will be connected in series and ligase acts like 'glue' to seal the covalent bonds between the adjacent fragments as shown in Figure 6 (Zucca, 2000).

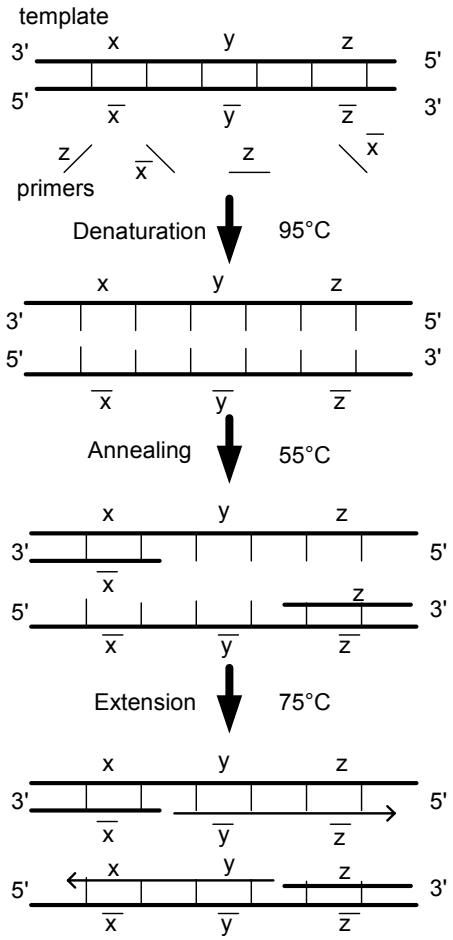

Figure 5. One cycle of PCR. For 30 cycles of PCR, these 3 steps are repeated for 30 times

4.6 Gel Electrophoresis

DNA strands in a solution can be separated in terms of its length by means of gel electrophoresis. The molecules are separated according to their weight, which is almost proportional to their length (Calude and Paun, 2001). This technique is based on the fact that DNA molecules are negatively charged (Paun *et al.,* 1998). Hence, by putting them in an electric field, they will move towards the positive electrode at different speeds. The longer molecules will remain behind the shorter ones.

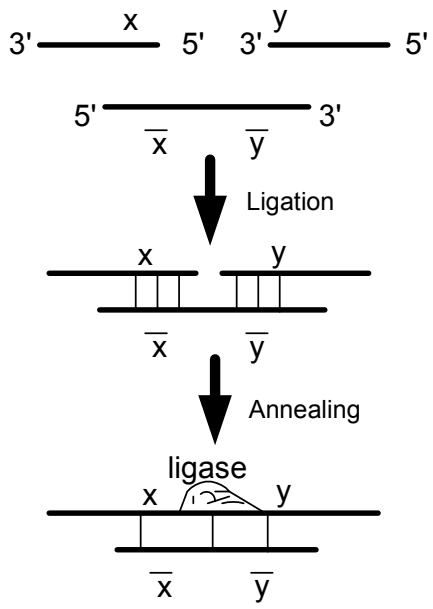

Figure 6. Example of the ligation process

The speed of the moving DNA molecules in a gel also depends on the gel porosity. Agarose gel is frequently used and by varying the porosity of the gel used, the sensitivity of this length separation operation can be altered. The precision is high even with molecules which only differ by one nucleotide. An example of gel electrophoresis process (Amos, 1997) and its output are well depicted in Figure 7 and Figure 8, respectively. This technique can be used to "print" the results of DNA computation as well. Normally, at the end of this process, the gel is photographed for convenience.

5. DNA SEQUENCE DESIGN AND SYNTHESIS BASED ON DIRECT-PROPORTIONAL LENGTH-BASED DNA COMPUTING

Consider a directed graph and the output of the shortest path computation as shown in Figure 3. Let n be the total number of nodes in the graph. The DNA sequences correspond to all nodes and its complements are designed. Let V_i (i= 1, 2, … , n) and $\overline{V_i}$ (i= 1, 2, … , n) be the 20-mer DNA sequences correspond to the ith node in the graph and its complement respectively. By using the available software for DNA sequence design, DNASequenceGenerator (Udo *et al.,* 1998), the DNA sequences V_i is designed and listed in Table 2. The GC contents (GC%), and melting temperature (Tm) of each sequence are also shown. In Table 2, V_i can be separated into V_{ia} and V_{ib} where V_{ia} is defined as half-5-end and V_{ib} is defined as half-3-end of V_i. Then, we introduce three rules to synthesize oligos for each edge in the graph as follows:

Rule (i) If there is a connection between V_1 to V_j, synthesize the oligo for edge as

$$V_1(20) + W_{1j}(\omega\text{- }30) + V_{ja}(10)$$

Rule (ii) If there is a connection between V_i to V_j, where $i \neq 1, j \neq n$, synthesize the oligo for edge as

$$V_{ib}(10) + W_{ij}(\omega\text{- }20) + V_{ja}(10)$$

Rule (iii) If there is a connection between V_i to V_n, synthesize the oligo for edge as

$$V_{ib}(10) + W_{in}(\omega\text{- }30) + V_n(20)$$

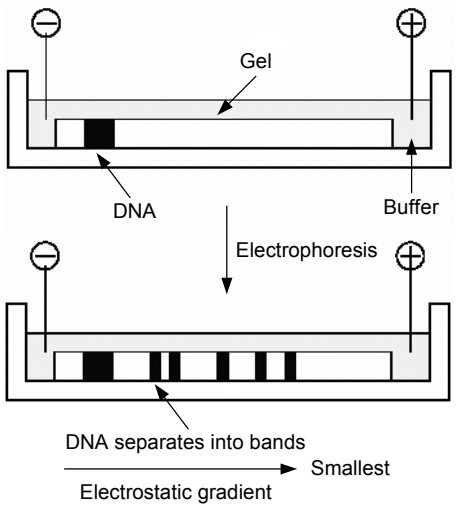

Figure 7. The gel electrophoresis process

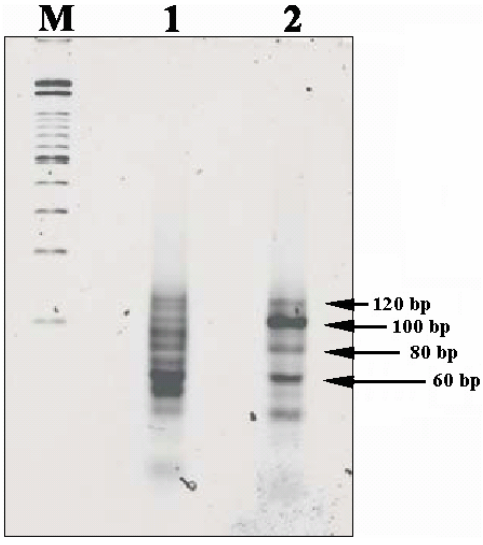

Figure 8. Gel electrophoresis output where lane M is the DNA size marker. Lanes 1

and 2 are used for the tested DNA molecules (Yamamoto *et al.*, 2002b)

Table 2. DNA sequences for nodes. The melting temperature, *Tm* are calculated based on Sugimoto (Sugimoto *et al.,* 1996) nearest neighbor thermodynamic parameter

Node, *Vi*	20-mer sequences (5'-3')		GC%	*Tm* (ºC)
	V_{ia}	V_{ib}		
V_1	TATGCTCATT	TGCATTTTGA	30	46.13
V_2	CAGGAGCGTC	TGAGAGCGAG	65	58.58
V_3	TACCGGATCG	ACCGGCTAAG	60	55.12
V_4	TCGTCCAAGG	GAGGCTTCTC	60	54.46
V_5	GCTATATTGC	GCTGGATGTG	50	52.16

Synthesized oligos consist of three segments. The number of DNA bases for each segment is shown in parenthesis and '+' represents the join. W_{ij} denotes the DNA sequences representing a cost between node V_i and V_j. The oligo is synthesized so that the number of DNA bases of that oligo and the cost at the corresponding edge are similar. As such, the concept of synthesis based on rule (i), rule (ii), and rule (iii) are depicted in Figure 9, Figure 10, and Figure 11, respectively. For an edge from V_i to V_j, due to the synthesized rules, the DNA sequences for distance or costs, W_{ij} are designed using the DNASequenceGenerator and the results are listed as shown in Table 3. Table 4 on the other hand lists all the synthesized oligos based on the proposed synthesis rules. As can be seen, after the synthesis, the number of DNA bases for the synthesized oligo and the costs at the corresponding edge are the same. The complement sequences of each node and cost are synthesized as well.

Table 3. DNA sequences for costs. Melting temperature, *Tm* are calculated based on Sugimoto (Sugimoto *et al.,* 1996) nearest neighbor thermodynamic parameter

Cost, *Wij*	DNA sequences (5'-3')	Length	GC%	*Tm* (ºC)
W_{13}	GAAGTGTACG TTAGGCTGCT	20	50	51.59
W_{45}	AAAGGTCGTC TTTGAACGAG	20	45	50.34
W_{34}	AAAGGCCCTC TTTTAACGAA GTCCTGTACT	30	43	60.45
W_{24}	AAAGCCCGTC GGTTAAGCAA GTAGTTTACG CTGCGTCATT	40	48	69.47
W_{25}	GCGTTGTTGC GAGGCATGTG GAGAATTGAT CGCTTTCGTG CATAACTGGG	50	52	74.08
W_{12}	CAGCATCGTA GTAGAGCTAG TATCGAACTG ATAAGTAACG GAGGGGGCTC	50	50	72.19
W_{23}	AAAGCTCGTC GTTTAAGGAA GTACGGTACT ATGCGTGATT TGGAGGTGGA	50	46	70.91

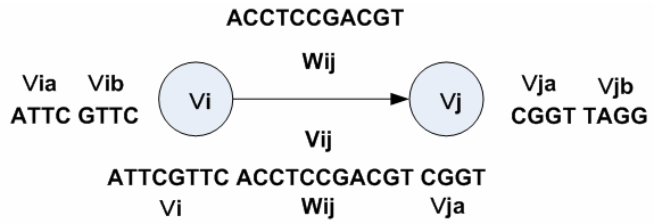

Figure 9. The concept of oligo synthesis for rule (i)

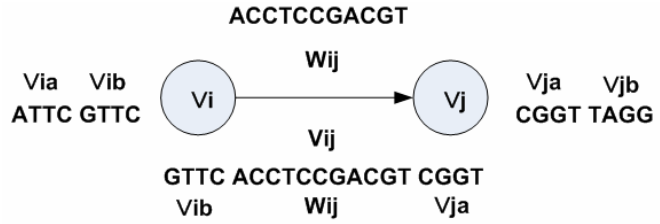

Figure 10. The concept of oligo synthesis for rule (ii)

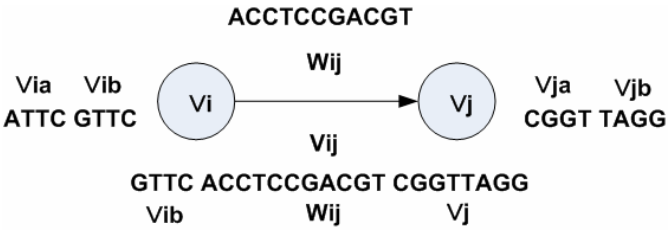

Figure 11. The concept of oligo synthesis for rule (iii)

Table 4. DNA sequences for edges

Edge	DNA sequences (5'-3')
$V_1 - W_{13} - V_{3a}$	TATGCTCATTTGCATTTTGAGAAGTGTACG TTAGGCTGCTTACCGGATCG
$V_{4b} - W_{45} - V_5$	GAGGCTTCTCAAAGGTCGTCTTTGAACGAG GCTATATTGCGCTGGATGTG
$V_{3b} - W_{34} - V_{4a}$	ACCGGCTAAGAAAGGCCCTCTTTTAACGAA GTCCTGTACTTCGTCCAAGG
$V_{2b} - W_{24} - V_{4a}$	TGAGAGCGAGAAAGCCCGTCGGTTAAGCAA GTAGTTTACGCTGCGTCATTTCGTCCAAGG
$V_{2b} - W_{23} - V_{3a}$	TGAGAGCGAGAAAGCTCGTCGTTTAAGGAA GTACGGTACTATGCGTGATTTGGAGGTGGA TACCGGATCG
$V_{2b} - W_{25} - V_5$	TGAGAGCGAGGCGTTGTTGCGAGGCATGTG GAGAATTGATCGCTTTCGTGCATAACTGGG GCTATATTGCGCTGGATGTG
$V_1 - W_{12} - V_{2a}$	TATGCTCATTTGCATTTTGACAGCATCGTA GTAGAGCTAGTATCGAACTGATAAGTAACG GAGGGGGCTCCAGGAGCGTC

6. COMPUTING WITH DNA

All the synthesized oligos are inserted into a test tube for initial pool generation. The generation of an initial pool solution is based on the hybridization/ligation method as depicted in Figure 12. In fact, the hybridization/ligation method for initial pool generation has been introduced firstly by Adleman (Adleman, 1994) in order to solve HPP. If the shortest path V_1 - V_3 - V_4 - V_5 is emphasized, Figure 13 clearly shows which kinds of oligos are important for the generation of this path. However, the unwanted combinations are also generated in the same manner.

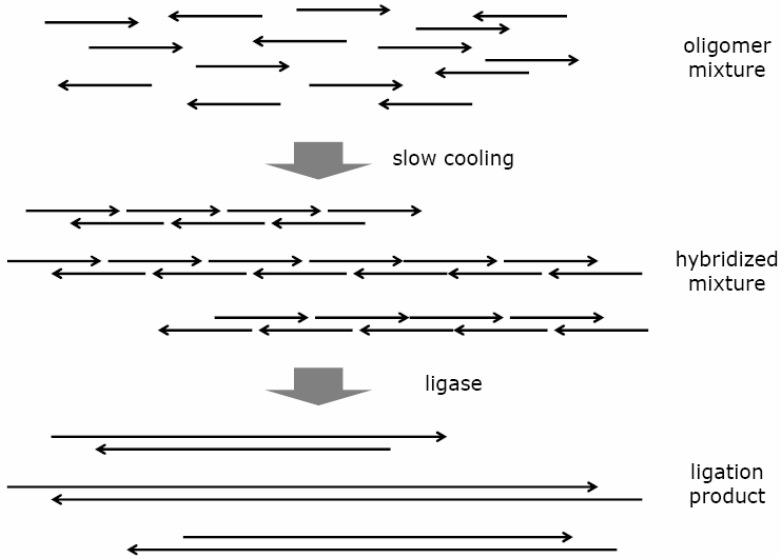

Figure 12. Hybridization/ligation method employed for initial pool generation. The arrowhead indicates the 3' end (Lee *et al.,* 2004)

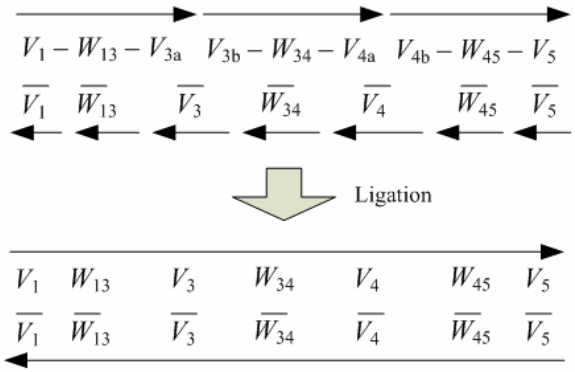

Figure 13. DNA duplex for path V_1 - V_3 - V_4 - V_5. The arrowhead indicates the 3' end

At this stage, an initial pool of solution has been produced and it is time to filter out the optimal combinations among the vast alternative combinations of the problem. Unlike conventional filtering, this process is not merely throwing away the unwanted DNA duplex but rather copying the target DNA duplex exponentially by using the

incredibly sensitive PCR process. This can be done by amplifying the DNA molecules that contain the start node V_1 and end node V_5 using primers. After the PCR operation is accomplished, there should be numerous number of DNA strands representing the start node V_1 and end node V_5 traveling through a possible number of nodes. Three types of possible DNA duplex amplified after the PCR process are given in Figure 14.

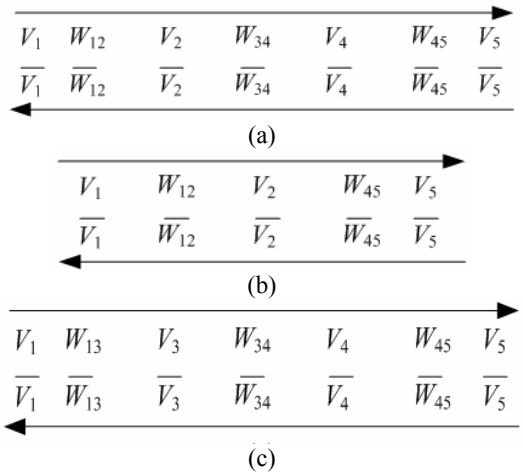

Figure 14. Examples of DNA duplex amplified by PCR. The length of DNA duplex in base-pairs (bp) is given in parenthesis and the arrowhead indicates the 3' end (a) DNA molecule $V_1 - V_2 - V_4 - V_5$ (190bp) (b) DNA molecule $V_1 - V_2 - V_5$ (160bp) (c) DNA molecule $V_1 - V_3 - V_4 - V_5$ (150bp)

The output solution of the PCR operation then undergoes gel electrophoresis operation. During this operation, the DNA molecules will be separated in terms of its length and hence, by analyzing the band of gel electrophoresis, the DNA duplex $V_1 - V_3 - V_4 - V_5$ representing the shortest path starting from V_1 and ending at V_5 would be extracted.

At this moment, based on the extracted shortest length DNA duplex, one only knows that the shortest path begins from V_1 and ends at V_5. However, this does not contain the information of the nodes that passed through the shortest path. The information regarding all the nodes in the shortest path can be obtained by applying the graduated PCR operation. In many cases, graduated PCR is used to allow one to "print" the result of the computation. For the sake of explanation, the DNA molecule representing the answer of the shortest path $V_1 - V_3 - V_4 - V_5$ is taken again for instance. Graduated PCR is performed by running four different PCR operations to the DNA strand $V_1 - V_3 - V_4 - V_5$ separately. The primer 5'–CTCGCTCTCAGACGCTCCTG– 3' is used to signal the node V_1 during the PCR operation. On the other hand, the primer 5'–CACATCCAGCGCAATATAGC–3', 5'–CTTAGCCGGTCGATCCGGTA–3', 5'–GAGAAGCCTCCCTTGGACGA–3', and 5'–CACATCCAGCGCAATATAGC–3' are employed as right primer at each separation to signal the node V_2, V_3, V_4, and V_5 respectively. It is expected that for the final solution containing the strand $V_1 - V_3 - V_4 - V_5$, 150 base-pairs (bp), graduated PCR will produce bands of x, 50, 100, and 150 in successive lanes of gel electrophoresis as depicted in Figure 15. This means that there is an intermediate node, V_3 and V_4 in between the start node V_1 and the end node V_5. The symbol x denotes the absence of a band corresponding to the omission of nodes V_2 along the

DNA duplex. Therefore, the shortest path of the graph can be readout as $V_1 \rightarrow V_3 \rightarrow V_4 \rightarrow V_5$.

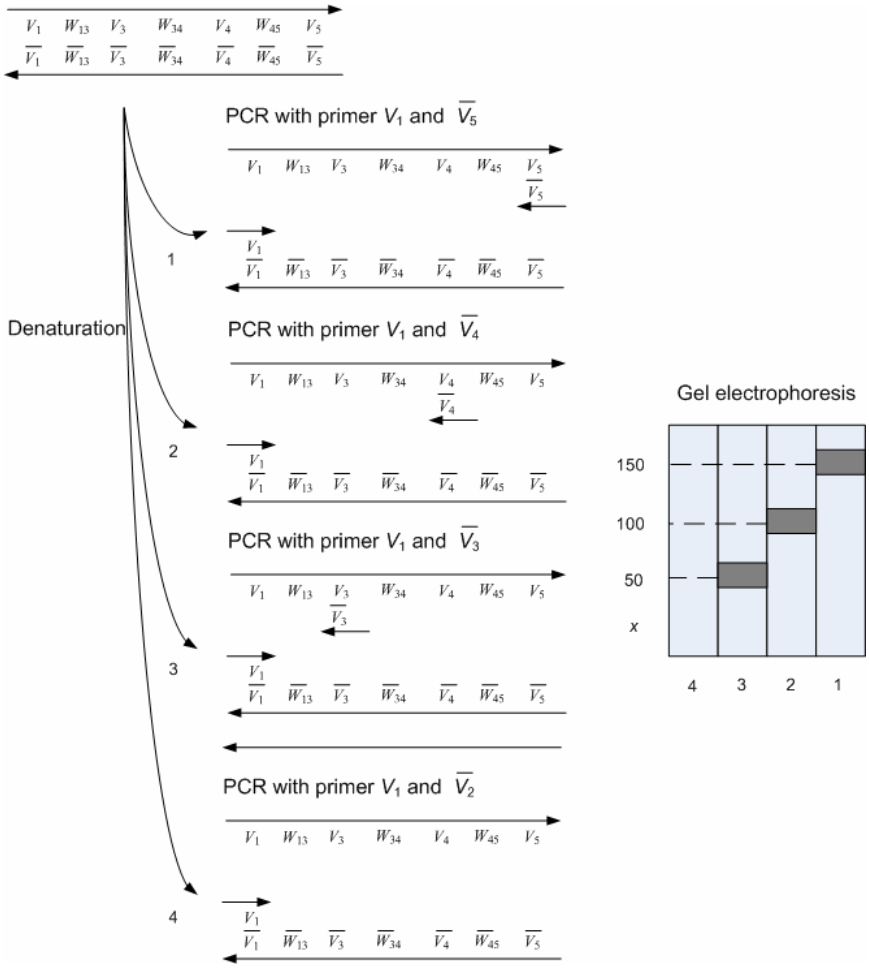

Figure 15. Example showing the results of the graduated PCR operation

5. CONCLUDING REMARKS

This paper has discussed a new approach called 'direct proportional length-based DNA computing' to solve weighted graph problems. In this approach, it is proposed that the direct proportional length of DNA is used to encode the cost of each edge. As the shortest path computation has been a benchmark problem in many applications, especially in engineering, it is expected that this approach can be easily extended to solve other types of weighted graph problems. Although the computational problem is encoded in a new way, all the laboratory operations involved during the computation are already well established in the field of DNA

computing. It is expected that the proposed approach, would extend the applicability of DNA computing in many types of applications in the future.

6. REFERENCES

Adleman L. (1994): *Molecular computation of solutions to combinatorial problems*. Science, Vol. 266, pp. 1021-1024.

Adleman L.M. (1998): *Computing with DNA*. Scientific American, pp. 34-41.

Amos M. (1997): *DNA computation*. PhD Thesis, The University of Warwick, UK.

Calude C.S. and Paun G. (2001): *Computing with cells and atoms: an introduction to quantum, DNA, and membrane computing.* New York: Taylor & Francis Inc.

Dijkstra E.W. (1959): A note in connexion with graphs. Numerische Mathematik, Vol. 1, pp. 269-271.

Feldkamp U., Saghafi S., Banzhaf W. and Rauhe H. (2002): DNASequenceGenerator: a program for the construction of DNA sequences. Lecture Notes in Computer Science, Vol. 2340, pp. 23-32.

Fitch J.P. (2002): *An engineering introduction to biotechnology*. Washington: SPIE-The International Society for Optical Engineering.

Lee J.Y., Lim H.W., Yoo S.I., Zhang B.T. and Park T.H. (2004): Efficient initial pool generation for weighted graph problems using parallel overlap assembly. Preliminary Proceedings of the Tenth International Meeting on DNA Based Computers, pp. 357-364.

Lee J.Y., Shin S.Y., Augh S.J., Park T.H. and Zhang B.T. (2002): Temperature gradient-based DNA computing for graph problems with weighted edges. Preliminary Proceedings of the Eighth International Meeting on DNA Based Computers, pp. 41-50.

Lee J.Y., Shin S.Y., Augh S.J., Park T.H. and Zhang B.T. (2003): *Temperature gradient-based DNA computing for graph problems with weighted edges*. Lecture Notes in Computer Science, Vol. 2568, pp. 73-84.

Moore G.E. (1965): *Cramming more components onto integrated circuits*. Electronics, Vol. 38, Number 8.

Narayanan A. and Zorbalas S. (1998): *DNA algorithms for computing shortest paths*. Proceedings of Genetic Programming, pp. 718-723.

Paun G., Rozenberg G. and Salooma A. (1998): *DNA computing: New computing paradigms*. Heidelberg: Springer-Verlag.

Sugimoto N., Nakano S., Yoneyama M. and Honda K. (1996): *Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes*. Nucleic Acid Research, Vol. 24, pp. 4501-4505.

Udo F., Sam S., Wolfgang B. and Hilmar R. (2001): *DNA sequence generator: A program for the construction of DNA sequences*. In N. Jonoska and N. C. Seeman (editors), Proceedings of the Seventh International Workshop on DNA Based Computers (DNA 7), pp. 23-32.

Yamamoto M., Kameda A., Matsuura N., Shiba T., Kawazoe Y. and Ahochi A. (2002): *A separation method for DNA computing based on concentration control*. New Generation Computing, Vol. 20, pp. 251-262.

Yamamoto Y., Kameda A., Matsuura N., Shiba T., Kawazoe Y. and Ahochi A. (2002): *Local search by concentration-controlled DNA computing*. International Journal of Computational Intelligence and Applications, Vol. 2, pp. 447-455.

Zucca M. (2000): *DNA based computational models*. PhD Thesis, Politecnico Di Torino, Italy.