Temperature Gradient-Based DNA Computing for Graph Problems with Weighted Edges

Ji Youn Lee[1], Soo-Yong Shin[2], Sirk June Augh[3],
Tai Hyun Park[1], and Byoung-Tak Zhang[2,3]

[1] Cell and Microbial Engineering Laboratory
School of Chemical Engineering,
[2] Biointelligence Laboratory
School of Computer Science and Engineering,
[3] Center for Bioinformation Technology (CBIT),
Seoul National University, Seoul 151-742, Korea
{jylee, syshin, sjaugh, thpark, btzhang}@bi.snu.ac.kr

Abstract. We propose an encoding method for numerical data in DNA using temperature gradient. To represent numerical values in DNA sequences, we introduce melting temperature (Tm). When DNA strands representing smaller values have a lower Tm, they tend to denature with ease and also easily amplified by denaturation temperature gradient polymerase chain reaction. We also implement a local search molecular algorithm using temperature gradient, which is contrasted to conventional exhaustive search molecular algorithms. The proposed methods are verified by solving an instance of the travelling salesman problem. We could effectively amplify the correct solutions and the use of temperature gradient made the detection of solutions easier.

1 Introduction

DNA computing has been applied to various fields of research. By the way, most of these applications do not consider the problem of representing numerical data in DNA molecules. However, many real world applications involve graph problems which have weighted edges. Examples include the shortest path problem, the travelling salesman problem, the minimum spanning tree problem, the Steiner tree problem [4].

We propose a novel encoding method that uses temperature gradient to solve the graph problems with weighted edges. The travelling salesman problem (TSP) is used as a benchmark problem for this technology. TSPs are interesting in that their solution requires to represent the path weights. This is contrasted with the Hamiltonian path problem [1] where the connection cost is binary. We represent the weights using melting temperature gradient, and implement a local search molecular algorithm with the temperature gradient method. Our bio-lab experiments show that the correct DNA strands are effectively amplified and the detection of optimal solutions made easier.

In Section 2, we describe the molecular encoding of numerical data in DNA molecules. Section 3 explains the molecular algorithm and experimental procedure for solving TSP. Section 4 presents the results and discusses them. The conclusions are drawn in Section 5.

2 Molecular Encoding for Edge Weights of Graph

2.1 Previous Encoding Methods

Narayanan and Zorbalas proposed DNA algorithms to solve travelling salesman problems [7]. Their method makes the length of weight sequences proportional to the edge costs in the edge sequences. This method is inefficient in representing a wide range of weights because, in some cases, one edge sequence with a high weight must be very long. For the same reason, it is also inefficient to represent real values. In addition, the fact that larger weights are encoded as longer sequences is contrary to the biological fact: the longer the sequences are, the more likely they hybridize with other DNA strands.

Yamamura et al. proposed a concentration control method to represent the weights [12]. The concentration of each DNA is used as input and output data, i.e. the numerical data is encoded by concentrations of DNAs. This method enables local search among all candidate solutions instead of exhaustive search. However, the concentration control method has some drawbacks in detecting the solutions. One cannot be sure that the most intensive band (in the gel) has the optimal solutions, because the most intensive band can be made by a number of partial non-optimal solutions with low concentrations rather than optimal solutions. In addition, it is technically difficult to extract a single optimal solution from the most intensive band.

In previous work [10], we proposed a method for representing real-values in fixed-length DNA strands using GC contents. Edge sequences contain two components: link sequences and weight sequences. The weight of edges is represented by varying GC content in weight sequences. Chemically, the hybridizations between the G/C pairs are preferred to those between A/T pairs. Therefore, the search procedure can be guided by including more A/T pairs to higher weight sequences and including more G/C pairs to smaller weight sequences.

2.2 Melting Temperature Control Encoding

Based on previous work [10], we improve the weight representation method by allowing the hybridization process to be controlled by melting temperature. We choose the melting temperature rather than the number of hydrogen bonds, since melting temperature (Tm) characterizes the stability of the DNA hybrid more exactly than GC content. Though the hybridization process is influenced by the number of hydrogen bonds between DNA pairs, the possibility of hybridization is much more dependent on Tm of given DNA sequences.

Many methods have been proposed to determine the Tm value of DNA duplex. In the absence of destabilizing agents, like formamide or urea, Tm depends

on three major parameters: the GC content and thermodynamic factors, the strand concentration, and the salt concentration. If we ignore the reaction conditions, Tm only depends on the GC content and thermodynamic factors. The classical method is using GC content, salt concentration, and the length [11]. But recently a statistical method using the thermodynamic parameters such as ΔH and ΔS is proposed [5]. The latter model is known as the nearest-neighbor (NN) model that is more accurate and applicable to DNA duplex up to 108 bp.

We use both the GC content method and the nearest-neighbor method to calculate the Tm. If the sequences are short, the NN method is used, otherwise the classical method is mainly applied. Though Tm of the DNA duplex up to 108 bp can be accurately determined by the NN method, Tm of the longer DNA strands formed after hybridization and ligation depends on the classical method rather than the NN method.

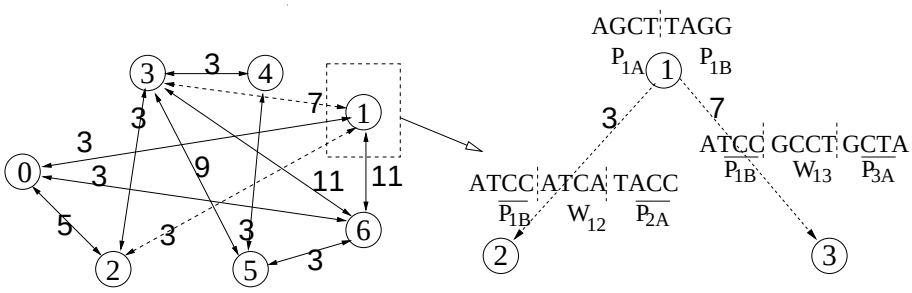

Fig. 1. Graph for the travelling salesman problem (left) and its encoding in DNA (right). The vertex sequence is $5' \rightarrow 3'$ and the edge sequence is $3' \rightarrow 5'$. The optimal path for this problem is '$0 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6 \rightarrow 0$'.

The melting temperature control encoding scheme is illustrated in Fig. 1. Basic representation is similar to Adleman's method [1]. The difference is that edge sequences have the weight sequence part in the middle of their sequences. First, the vertex sequences are designed. Each vertex sequence is designed to have a similar Tm using both the nearest-neighbor and the classical method. That is why vertex sequences should not affect the hybridization fidelity. Then, the edge sequences are generated based on the vertex sequences to be linked. The edge sequence consists of the two parts such as the link sequence and the weight sequence. This scheme is motivated from [8]. The link sequence has also a similar Tm, since the link sequence of the edge sequence is decided by the vertex sequence. The weight sequence is designed to have a various Tm according to its representative cost. To produce more variation in generating low cost paths, the weight sequences with smaller weights have a lower Tm. If DNA strands have a lower Tm, it can be easily denatured and also easily amplified by denaturation temperature gradient polymerase chain reaction. The Tm of weight sequences is decided by the classical method, because the partial paths are long enough

Table 1. Vertex sequences and weight sequences for TSP. Tm is calculated by the NN method.

Vertex sequences			
No.	Sequence ($5' \rightarrow 3'$)	Tm	GC%
0	AGGCGAGTATGGGGTATATC	60.73	50
1	CCTGTCAACATTGACGCTCA	59.24	50
2	TTATGATTCCACTGGCGCTC	59.00	50
3	ATCGTACTCATGGTCCCTAC	56.81	50
4	CGCTCCATCCTTGATCGTTT	58.13	50
5	CTTCGCTGCTGATAACCTCA	59.44	50
6	GAGTTAGATGTCACGTCACG	56.97	50
Weight sequences			
Edge cost	Sequence ($5' \rightarrow 3'$)	Tm	GC%
3	ATGATAGATATGTAGATTCC	47.89	30
5	GGATGTGATATCGTTCTTGT	54.62	40
7	GGATTAGCAGTGCCTCAGTT	58.37	50
9	TGGCCACGAAGCCTTCCGTT	64.51	60
11	GAGCTGGCTCCTCATCGCGC	68.88	70

not to fit the NN model. For example, as shown in Fig. 1, the first part ($\overline{P_{1B}}$) of the edge sequence is complementary to the last half (P_{1A}) of the starting vertex sequence. And the last part ($\overline{P_{2A}}$) is complementary to the first half of the ending vertex sequence (P_{2A}). The weight sequence W_{12} with cost 3 contains more A/T pairs than the weight sequence W_{13} with cost 7 to vary the Tm.

2.3 Travelling Salesman Problems

We selected travelling salesman problem as a benchmark. The travelling salesman problem (TSP) is to find a minimum weight (cost) path for a given set of vertices (cities) and edges (roads). In addition, the solution path must contain all the cities given, each only once, and begin from the specified city to which the tour ends [4]. We solve the TSP with 7 nodes, 23 edges and 5 weights as shown in Fig. 1. For convenience, we represent the weights by decimal numbers such as 3, 5, 7, 9, and 11. The example of DNA sequences for solving TSP is shown in Table 1. These sequences are generated by the NACST sequence generator (NACST/Seq) [6]. The detailed constraints for the sequences design are shown in [9]. For the vertex sequences, Tm ranges 55^oC to 60^oC using the NN model with $1M$ salt concentration and $10nM$ oligomer concentration, GC content is 50%, the unexpected secondary structure such as hairpin formation is prohibited. In the weight sequences, GC content is varied from 30% to 70% (see Table 1). While weight sequences do not affect the hybridization process (partial solution generation step), they are important in detection when the sequences are long enough not to fit the NN model. Therefore we varied GC contents of weight sequences. To avoid the extreme case, i.e. the sequence consists of all C or G, we set the range as $30\% \sim 70\%$. As the result of evolutionary algorithms, NACST/Seq assigns the 30% to weight 3, 40% to weight 5, and so on. NACST/Seq increases GC content by 10%, as the weight is increased by two. Additionally, the higher weight sequences have higher melting temperatures with the NN method.

3 Molecular Algorithms for TSP

3.1 Previous Work

Most of the existing DNA computing methods to solve combinatorial optimization problems are similar to Adleman's in several ways. First, all possible solutions are generated and then the correct solutions among them are selected. Also the previous works [7, 10, 12] on travelling salesman problems or shortest path problems are similar to each other. They only differ in their weight encoding schemes such as sequence length, GC content, and strand concentration. The former two studies were not verified by lab experiments, the last method is proved by lab experiments, though the results are not satisfactory. And last method [12] gives a bias to search space by concentration control.

3.2 Temperature Gradient Algorithms

In order to utilize the massive parallelism of biochemical materials and to guide search space, it is useful to give some biases in the experimental steps. For this purpose, we give some constraints in DNA concentration and the PCR step. By varying the concentration of each oligomer strands [12], we can generate more paths that contains a smaller sum of weights. By modifying the PCR protocol, we can specifically amplify and detect the correct solution easier.

DNA strands of low melting temperature can be more strongly amplified in PCR by varying the denaturation protocol. Usually, we vary the annealing temperature to optimize the reaction condition. But by denaturing the dsDNA in the lower temperature in the starting cycles and gradually increasing the denaturation temperature, we can specifically amplify the DNA strands with a lower melting temperature. So we can amplify the correct solutions more than other solutions that have the same length but a higher melting temperature. The whole procedure is shown in Fig. 2. In step 1, we generate the solution set. Steps $2 \sim 5$ find the solution satisfying the restrictions (the starting vertex is the same as the end vertex). Step 6 separates the sequences visiting all vertexes. In steps $7 \sim 8$, we amplify the solutions based on its path cost and find the optimal solution. Finally, we check the path of the solution.

Oligomer synthesis: All 7 vertexes and 5 weights are designed in 20 mer ssDNA as listed in Table 1. And 23 edges in 40 mer ssDNAs are designed according to their vertexes and weight sequences. All oligomers are 5'-phosphorylated. Other 5'-biotinylated vertexes were prepared for affinity-separation. All oligomers were synthesized at Bioneer Co.

Hybridization and ligation: Edge and weight oligomers were differently added by their weights. The amount decreased by 20% according to the weight increase by 2, set on the basis of weight 3 as a 100%. Oligomer mixture was heated to 95^oC and cooled to 20^oC by 1^oC per 1 minute, and stored at 4^oC.

Mix $5\mu l$ of hybridization mixtures and 350 units of T4 DNA ligase (TaKaRa, Japan), ligase buffer ($66mM$ Tris-HCl, pH 7.6, $6.6mM$ $MgCl_2$, $10mM$ DTT,

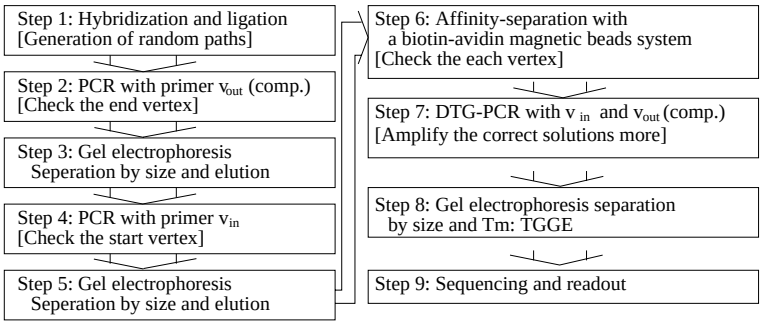

Fig. 2. The molecular algorithms for solving TSP.

$0.1mM$ ATP), and add H_2O to a total volume of $10\mu l$. We incubated the reaction mixture for 16 hours at 16^oC.

PCR amplification and gel electrophoresis: All PCR amplifications were performed on a PTC-200 DNA Engine (MJ Research, MA, USA). For normal amplification, $1\mu M$ of primer and AccuPower PCR PreMix (Bioneer, Korea) which containing 1 unit of Taq DNA polymerase in $10mM$ $Tris$-HCl, pH 9.0, $1.5mM$ $MgCl_2$, $40mM$ KCl, $0.25mM$ each dNTP were dissolved to distilled water to a total volume of $20\mu l$. And PCR was processed for 34 cycles at 95^oC for 30 seconds, at $(Tm-5)^oC$ for 30 seconds and at 72^oC for 30 seconds. Initial denaturation and prolonged polymerization was executed for 5 minutes.

All gel electrophoresis were performed with 2% Agarose-1000 (GibcoBRL, NY, USA) in 0.5X tris-borate-EDTA buffer and the gel was ethidium bromide stained. 50bp DNA Ladder (GibcoBRL, NY, USA) was used as a marker.

Affinity separation: We roughly followed the affinity separation protocol in the Adleman's experiments. The ssDNA for the affinity purification were produced by replacing the forward primer with 5'-biotinylated analog. The amplified product was annealed to streptavidin paramagnetic particles (Promega, Madison, WI) by incubating in $200ml$ of $0.5\times$ saline sodium citrate (SSC) for 1 hour at room temperature with constant shaking. Particles were washed three times in $300ml$ of $0.5\times$ SSC and then heated to 80^oC in $100ml$ of ddH_2O for 5 minutes to denature the bound dsDNA. The aqueous phase with ssDNA was retained. For affinity separation, $1n$mol of 5'-biotinylated vertex strands was annealed to particles as above and washed three times in $300ml$ of $0.5\times$ SSC for 1 hour at room temperature with constant shaking. Particles were washed four times in $300ml$ of $0.5\times$ SSC to remove unbound ssDNA and then heated to 80^oC in $100ml$ of ddH_2O for 5 minutes to release ssDNA bound to the complementary vertex 1. The aqueous phase with ssDNA was retained. This process was then repeated with each vertexes.

Denaturation temperature gradient PCR: For denaturation temperature gradient PCR (DTG-PCR), the denaturation temperature was kept initially 70^oC and gradually decreased 1^oC per one cycle and kept 95^oC for the remaining

10 cycles. Other conditions are identical with normal PCR described in the PCR amplification section.

4 Results and Discussion

4.1 Random Path Formation and Size Sieving

Though searching a correct solution in the candidate pool is an important step to solve the problem, generating the more possible solution is also important in molecular computing. The reaction rate of biochemical reaction is related with the reaction constant and the reactant concentrations. Therefore, concentration gradient can be useful to obtain the correct solution easily. This approach was already tried by [12]. By increasing the concentration of DNA sequences of smaller weights and vice versa, paths which contains a smaller sum of weights will be more frequently generated. In our work, the concentration gradient was only a tool to generate more possible solutions.

The hybridization and ligation results are shown in Fig. 3 (A). Compared with the oligomer mixture, the ligated DNA strands became elongated. But, there are few copies around the 300 bp that indicates the length of paths including 8 vertexes. So we executed the ligation reaction again and obtained an upper shifted ligation product as shown in Fig. 3 (B). However, still shorter DNA strands mainly occupy the ligation product. Ligation is an essential step to generate random paths, so efficient ligation is needed to produce DNA strands long enough to solve large problems.

4.2 PCR with In and Out Vertex and Affinity Purification

After the hybridization and ligation reaction, dsDNAs with sticky ends are generated. We cannot use the primer pair because the start and the end point are identical in TSP and the primer pair is exactly complementary to each other. So we executed PCR with only one primer that is complementary with the end vertex, 0. By this PCR, we could make blunt-ended double-stranded DNAs that end with the vertex 0. We sieved the PCR product by 2% agarose gel electrophoresis (Fig. 3 (C)) and picked out the DNA strands around 300 bp. Subsequently, we executed the second PCR with vertex 0, and we could amplify the DNAs that start with vertex 0. As shown in Fig. 3 (D), the PCR product appeared in several bands in the gel. We excised and eluted the band around 300 bp and amplified with 5'-biotinylated vertex 0 as a primer. The amplified product was used in the following affinity separation with streptavidin paramagnetic particles.

4.3 Denaturation Temperature Gradient PCR and Latter Implementation Steps

Usually, we denature dsDNAs to ssDNAs by heating the mixture to 95°C. We made the Tm of the correct solution to be the lowest among the candidate solutions using the temperature gradient method. Therefore, if we decrease the

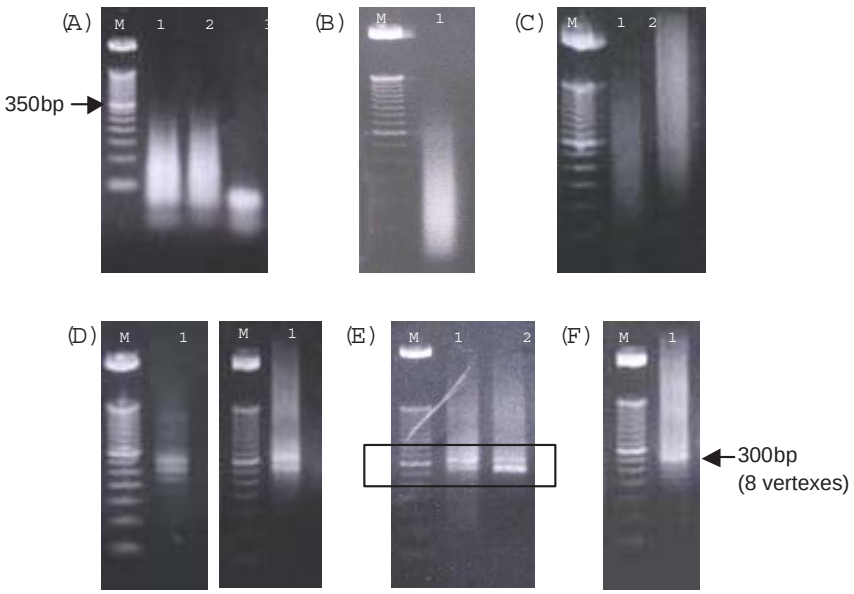

Fig. 3. Gel electrophoresis on a 2% agarose gel. The lanes contain: lane M denotes DNA size marker (50 bp ladder), (A) lanes 1,2: first ligation result, lane 3: oligomer mixture, (B) lane 1: second ligation result, (C) lanes 1,2: first PCR with V_{out} result, (D) lane 1: second PCR with V_{in} result, (E) lane 1: normal PCR result, lane 2: DTG-PCR result, (F) lane 1: the final DTG-PCR result.

denaturation temperature to a certain level, mainly DNA strands of correct solutions can be denatured at that temperature, and be amplified. As the denaturation temperature increases, other DNA strands also will be amplified. But the amount of correct solutions will be more increased cycle by cycle, and occupy the major part of the solution.

By simple modification of the typical PCR, we can amplify the correct solution and detect it easily. We can show the effectiveness of DTG-PCR by the relative amplification in Fig. 3 (E). Shorter DNA strands in the lane 2 are more amplified by DTG-PCR when compared with the normal PCR product in the lane 1. We purified the lower band by excision and elution from the gel, and executed the repetitive DTG-PCR, and obtained the band around 300 bp. This band contains four different DNA strands of the possible Hamiltonian paths, '$0 \to 1 \to 2 \to 3 \to 4 \to 5 \to 6 \to 0$ (the sum of weights: 21)', '$0 \to 1 \to 6 \to 5 \to 4 \to 3 \to 2 \to 0$ (31)', '$0 \to 2 \to 1 \to 3 \to 4 \to 5 \to 6 \to 0$ (27)', '$0 \to 2 \to 3 \to 4 \to 5 \to 6 \to 1 \to 0$ (31)'. Because these DNA strands have the identical length, 300 bp, they cannot be separated by normal gel electrophoresis. However, we can separate these strands by the temperature gradient gel electrophoresis (TGGE), because they have distinct melting behaviors. Similar work has studied using denaturing gradient gel electrophoresis (DGGE) [3].

TGGE that uses temperature as a denaturing agent is a modified form of DGGE. The GC content of the strands varies from 40.67% to 44.00%, and the melting temperature of the DNA strands are respectively 84.80^oC, 86.51^oC, 85.68^oC, 86.62^oC by NN model [5] with $1M$ salt concentration and $10nM$ oligomer concentration, and 92.69^oC, 94.05^oC, 93.51^oC, 94.05^oC by GC content method. The separated DNA strands with the lowest Tm can be sequenced to confirm that the whole procedure is correct. If we design the starting and end vertex to contain certain restriction enzyme sites, the sequencing step will become easier. These latter implementation steps will be a future work.

4.4 Scalability

The solvable problem size is restricted to the capabilities of the current biochemical tools. In case of solving TSP, PCR is the critical step in the procedure. We can amplify DNA strand of up to 40,000 bases with current PCR technology. If we apply this temperature gradient encoding method, we can scale up to 1,000 city problems. The crucial point of scalability is the ability of separation and detection. We can separate DNA strands whose GC content differs by 0.49 % with the current resolution of TGGE and this difference is easily achievable with our encoding method.

However, there exist some limitations in scaling up to 1,000 city problems. The first and critical problem is that it is impossible to discriminate the solutions which have the identical melting temperatures. And it is not also discriminative when there exist two solutions that melting temperature differences are not exceeding 0.2^oC. The second problem is the presence of elaborate and time-consuming implementation steps. In the experimental view, 1,000 times affinity separation is almost impossible. This is the most time-consuming step in our implementation steps. But, if we introduce the affinity separation method proposed by [2], the affinity separation steps can be automated with high efficiency and can became easier. Other possible obstacle in our encoding method arises when the DNA strands with stable mismatches have lower melting temperature and are amplified exclusively. So the sophisticated sequence design to avoid the stable mismatches is needed.

5 Conclusions

We introduced a temperature gradient method to solve the problems with numerical data. This method can overcome the restrictions of other encoding methods and be easily implemented by a simple modification of experiments. We can more strongly amplify the correct solution relatively during the reaction and the solution can be easily detected. This method drives the DNA pool to contain more correct solutions, rather than to search randomly for the correct solution in the DNA pool.

We also showed the feasibility of our computing method by solving the travelling salesman problem with adequate biochemical tools. The combination of

the orthogonal design of DNA sequences and denaturation temperature gradient PCR provides a novel method to solve general graph problems with weighted edges. This is applicable to any Tm-involved implementation steps to amplify low melting temperature DNA strands.

Acknowledgments

This research was supported in part by the Ministry of Education & Human Resources Development under the BK21-IT Program and the Ministry of Commerce, Industry and Energy through the Molecular Evolutionary Computing (MEC) Project. The RIACT at Seoul National University provides research facilities for this study.

References

1. L. M. Adleman. Molecular computation of solutions to combinatorial problems. *Science*, 266:1021–1024, 1994.
2. R. S. Braich, N. Chelyapov, C. Johnson, P. W. K. Rothemund, and L. Adleman. Solution of a 20-varaible 3-SAT problem on a DNA computer. *Science*, 296:499–502, 2002.
3. J. Chen, E. Antipov, B. Lemieux, W. Cedeño, and D. H. Wood. *In vitro* selection for a MAX 1s DNA genetic algorithm. In *Proceedings 5th International Workshop on DNA-Based Computers*, pages 23–46, 2000.
4. M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. W. H. Freeman and company, 1979.
5. J. SantaLucia Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA*, 95:1460–1465, 1998.
6. D. Kim, S.-Y. Shin, I.-H. Lee, and B.-T. Zhang. NACST/Seq: A sequence design system with multiobjective optimization. *8th International Meeting on DNA Based Computers* (Accepted).
7. A. Narayanan and S. Zorbalas. DNA algorithms for computing shortest paths. In *Proceedings of Genetic Programming 1998*, pages 718–723, 1998.
8. Q. Ouyang, P. D. Kaplan, S. Liu, and A. Libchaber. DNA solution of the maximal clique problem. *Science*, 278:446–449, 1997.
9. S.-Y. Shin, D.-M. Kim, I.-H. Lee, and B.-T. Zhang. Evolutionary sequence generation for reliable DNA computing. *Congress on Evolutionary Computation 2002*, (Accepted).
10. S.-Y. Shin, B.-T. Zhang, and S.-S. Jun. Solving traveling salesman problems using molecular programming. In *Proceedings of Congerss on Evolutionary Computation 1999*, pages 994–1000, 1999.
11. J. G. Wetmur. DNA probes: applications of the principles of nucleic acid hybridization. *Crit. Rev. Biochem. Mol. Biol.*, 26:227–259, 1991.
12. M. Yamamura, Y. Hiroto, and T. Matoba. Solutions of shortest path problems by concentration control. In *Proceedings of 7th International Workshop on DNA-Based Computers*, pages 231–240, 2001.