International Journal of Computer Science & Applications
Vol. II, No. II, pp. 118 - 130

DNA Computing Approach to Semantic Model

Yusei Tsuboi, Zuwairie Ibrahim, and Osamu Ono

Control Systems Laboratory
Institute of Applied DNA Computing
Graduate School of Science & Technology
Meiji University
1-1-1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa, 214-8671, JAPAN
e-mail: {tsuboi, zuwairie, ono}@isc.meiji.ac.jp

Abstract:

In this paper, we propose a new semantic model based on DNA computing. In the model the vertexes denote either a name of the target or both the attributes and attribute values. One path from an initial vertex to a terminal vertex means one object named on the tag. We name this model a *semantic model based on DNA computing*. The model explains a target object is reasoned out by the combinations between the vertexes. We describe its application to reasoning system by using DNA computing algorithm from theoretical point of a view. Vertexes and edges of the model are encoded to four kinds of nucleotides. The model is represented by double-stranded DNAs. Single-stranded DNAs are hybridized and ligated to let them the double-stranded DNAs, with the complementary sequences of input molecules and knowledge based ones. The generated DNAs are analyzed into necessary strands which are double-stranded DNAs representing the target objects. We discuss the proposed model and application with a simulation of computational complexity.

Keywords: DNA computing, semantic net, knowledge base, reasoning system, algorithm

1. INTRODUCTION

In DNA computing, information is stored in molecules that are linear polymers composed of nucleotides. Adleman's [1] ground-breaking work demonstrated the way to use molecules for computational purposes. The extreme compactness of DNA as data storage is nothing short of incredible. Reif [2] reported these characteristics of DNA computing as follows. Since a mole contains 6.02×10^{23} DNA base monomers, and the mean molecular weight of a monomer is approximately 350 grams/mole, then 1 gram of DNA contains 1.7×10^{21} DNA based. Since there are 4 DNA bases can encode 2 bits, and it follows that 1 gram of DNA can store approximately 3.4×10^{21} bits. In contrast,

conventional storage technologies can sore at most roughly 10^9 bits per gram, then DNA has the potential of storing data on the order of 10^{12} more compactly than conventional storage technologies. Baum [3] first proposed the idea of using DNA annealing to do parallel associative search in large databases encoded as DNA strands. The idea is very appealing since it represents a natural way to execute a computational task in massively parallel fashion. Moreover, the required volume scales only linearly with the base size. Retrievals and deletions under stringent conditions occur reliably (98%) within very short times (100 milliseconds), regardless of degree of stringency of the recall or the number of simultaneous queries in the input. Arita [4] suggests encoded data and report his experimental results of performing concatenation and rotation of DNA. His work also shows the possibility of join operations in relational database with molecules.

However, these models regarding the associative memory or database are not based on knowledge representation for artificial intelligence. It is thought that one method of approaching a memory with power near to that of man is to construct the semantic model based on molecular computing. It is important to propose a new model such as the above suitable to newtype-computers like DNA-based computers for further development of artificial intelligence.

Our research group has been focusing on developing a semantic net (semantic network) area by using a new computational paradigm. In Quillian's model [5-6] of semantic memory, concepts are represented by the name of relationship by links. Links are labeled by the name of relationship, and are assigned "criteriality tags" that attest to the importance of link. In artificial computer implementations, criteriality tags are numerical values the represent degree of association of the two concepts (such as how often that link is traversed) and the nature of the association. What type of the model is most effective to DNA-based computers which have a lot of potentials? In this paper, we proposed a new semantic model and its application to reasoning system by using DNA computing algorithm. In addition, we evaluate the proposed application with a simulation of computational complexity.

2. DESCRIPTION OF A SEMANTIC MODEL BASED ON DNA COMPUTING

Here, we describe how to design a semantic model based on DNA computing. Semantic net would mimic intellectual ability of people if they were based on associative memories. Its structure is a two-dimensional graph like a network. It is relatively easy for a man to deal with semantic net, because it represents an object (or concept) constructed from knowledge based on human's memories. The semantic net is made of three relations, an object, O; an Attribute, A; an Attribute Value, V. In general, these list representations are denoted as follows,

$$\{<O, A_i, V_{ji}> \mid i=1, 2,\ldots,m; j=1,2,\ldots,n\} \quad (1)$$

A basic semantic net as a graph is designed with vertexes, directed edges, and label representing their relations, as shown in Figure 1. Term O is reasoned out by relation between A_i and V_{ji}. Because the semantic net is simply standardized with vertexes and edge between them, it is suitable for a system to search for some objects in parallel.

Vertex ⇔ Object and attribute value
Directive Edge ⇔ Attribute

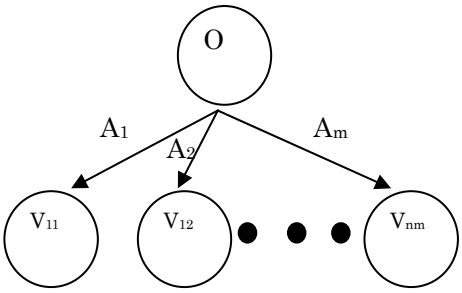


Figure 1: A basic Semantic net

The other hand, if there is a complicated graph, it is imperative to transfer it into simple one. AND/OR graph enables us to lessen the size of the graph, and to understand it more easily. We demonstrate a new model using this graph and describe the way to design it. Thus, we do not employ the normal existent semantic net like Figure 1, but the new model defined by us in order to make the most of DNA computing.

First, a tag as a name of a target object is set to an initial vertex in the graph. After we determine the number and the kinds of the attributes of the target object, both the attributes and attribute values are also set to another vertex following by the tag vertex. Second, the relation between vertexes and edges is represented using a new defined AND/OR graph. In Figure 2-a a directive edge in the terminal direction is connected between the vertexes in series except for the following case. If there are two vertexes which have same attributes but different attribute values, each of directive edges is connected in parallel as shown in Figure 2-b. The edge denotes only connection between the vertexes in the directive graph. Finally, labels are attached to the vertexes, such as '(Tag)' and '(Attribute, Attribute Value)'.

The vertexes denote either a name of the target or both the attribute and attribute value. In short, one path from an initial vertex to a terminal vertex means one object named on the tag. We define this model as a *semantic model based on DNA computing (SMD)*. The model explains a target object is reasoned by the combinations between the vertexes. For example, we design the model in the case of an apple named on the tag as shown in Figure 3.

The limitation of the knowledge representation ability depends on existent semantic nets in many cases because the SMD is derived from the semantic nets. For instance, if SMD uses dynamic knowledge, some special engines would be required to update it. However, under now physical equipment its ability is very higher than the other models [3-4][7-8] since it is composed of both attributes and attribute values with a concept of AND/OR graph.

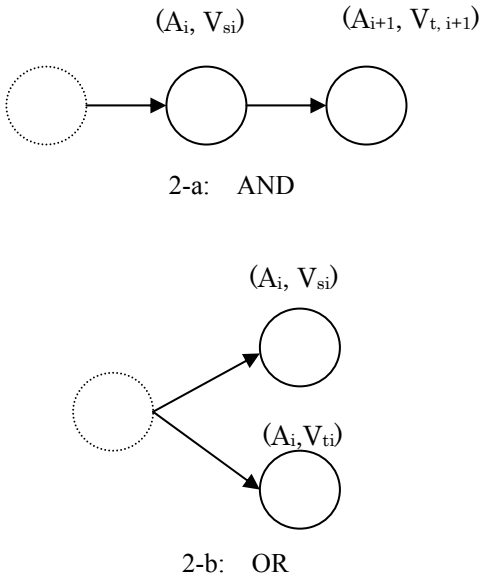


Figure 2: AND/OR graph

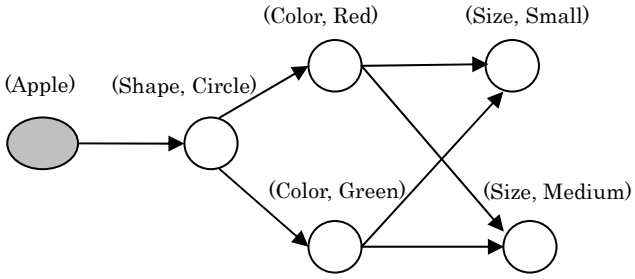


Figure 3: In apple case, semantic model based on DNA computing (SMD)

3. APPLICATION TO REASONING SYSTEM BASED ON SMD

In this section, application to a reasoning system based on SMD is demonstrated in order to illustrate effectiveness of SMD. We outline necessary chemical processes for the application from theoretical point of a view.

3.1 Sequence Design

In SMD, each of vertexes and edges is represented by a DNA strand as follows: the vertexes, except the tag's one, are represented by a short (0-12 nucleotide) piece of a single-stranded sequence by Table 1. In Table 1, a row shows attributes, a column shows attribute values and DNA sequence is randomly designed by these relations so that it might not be overlapped with the other sequences. Every tag vertex of the objects has random sequences of unique number (200, 300, 400…) at the left end to distinguish the objects, as shown in Figure 4. The sequences of the edge $(A_i, V_{si})\rightarrow(A_{i+1}, V_s)$ from vertex (A_i, V_{si}) to vertex $(A_{i+1}, V_{t,\ i+1})$ are complementary to the vertex sequences derived from the 3' 6- mer of vertex (A_i, V_{si}) and from the 5' 6-mer of vertex$(A_{i+1}, V_{t,\ i+1})$, as shown in Figure 5. It is a short (0-12 nucleotide) piece of single-stranded DNA except initial and terminal edges. These two DNA pieces are represented respectively by the size which suits the end of the DNA pieces of the initial or the terminal exactly.

In this way, the SMD is represented by double-stranded DNAs. Figure 6 shows one ((Apple)→(Shape, Circle)→(Color, Red)→(Size, Medium)) of the paths in the case of the apple is represented by a double-stranded DNA.

Table 1: Design of attributes and attribute values (vertex)

Shape		Color		·····		Size	
Circle	TCGATCTACTTA	White	TACGATTCGGAT	·····	·····	Large	TACTCGATACAT
Square	TTGCATCGTTAC	Red	ATCGTACCTGAT	·····	·····	Medium	CAGCTGAAATCA
				·····	·····		
Triangle	ATCCATGGATCG	Green	GTTATTCCCCAG	·····	·····	Small	AATTACGGGATA

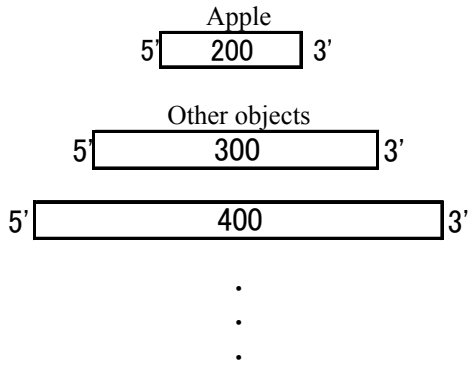


Figure 4: Tag design

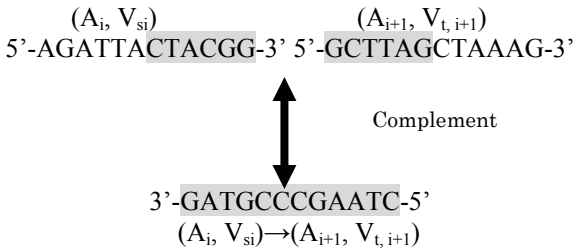


Figure 5: Edge design

Path : (Apple)→(Shape, Circle)→(Color, Red)→(Size, Medium)

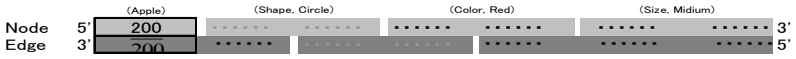


Figure 6: Double- stranded DNA (Apple)

3.2 Reaction

The one to P reference objects are reasoned out by using the DNA computing algorithm. We outline defined processes of the chemical reaction for the solution by theoretical virtual operations. Figure 7 shows the reasoning system composed of the knowledge base and a premise for the solution. In the knowledge base, each of DNA pieces of edges and tags is synthesized as a *knowledge based molecule*. In the premise, the attribute values are extracted from 1-P reference objects separately under the previously determined attributes. These labels are represented as '(Attribute, Attribute Value)'. Using the (Attribute, Attribute Value), a piece of a single-stranded DNA is synthesized as an *input molecule* by Table.1.

We have to amplify each of the knowledge based molecules and the input molecules sufficiently with PCR (Polymerase Chain Reaction). PCR is a technique which is used to amplify the number of copies of a specific region of DNA, in order to produce enough DNA to be adequately tested. This technique can be used to identify with a very high-probability, disease-causing viruses and/or bacteria, a deceased person, or a criminal suspect.

Figure 8 shows DNA operations for treating input molecules and knowledge based molecules. Now, we prepare virtual 1-P tubes to reason out 1-P reference objects. The knowledge based molecules divided into 1-P equal aliquots are put into each test tube. Each of 1-P input molecules are put into 1 to P test tubes respectively. Figure 9 shows the single-stranded DNAs will anneal to a complementary sequence under defined reaction conditions in each test tube. The DNA sequences representing input molecules and knowledge based molecules are mixed in the presence of a DNA ligase. This enzyme will form a covalent bond between two DNA molecules as long as they have complementary single-stranded overhangs. Thus, the sequences are ligated to form a duplex DNA which represents a path between an initial vertex and a terminal vertex. As a result, all the possible double-stranded DNAs representing the paths are generated at random.

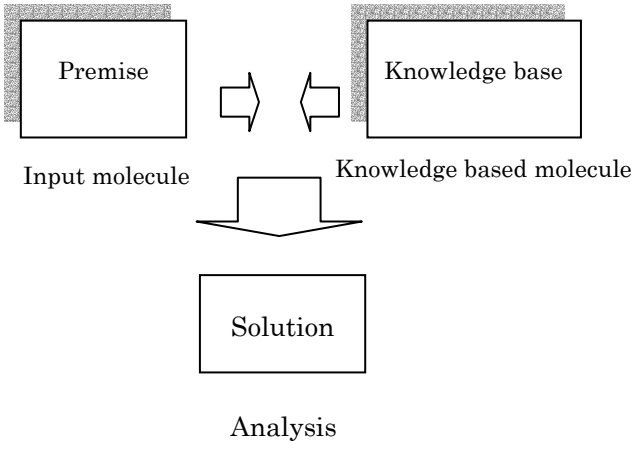


Figure 7: Flowchart of the reasoning system

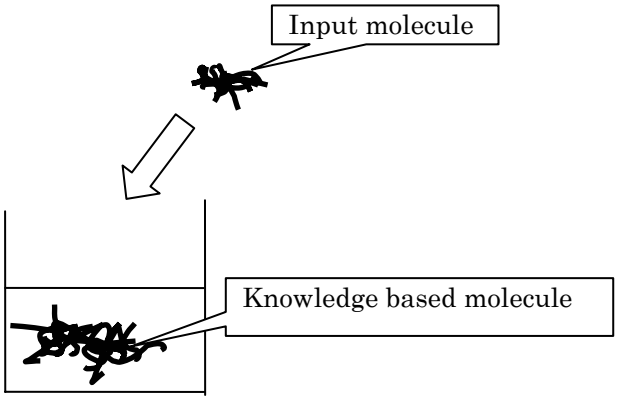


Figure 8: DNA operations

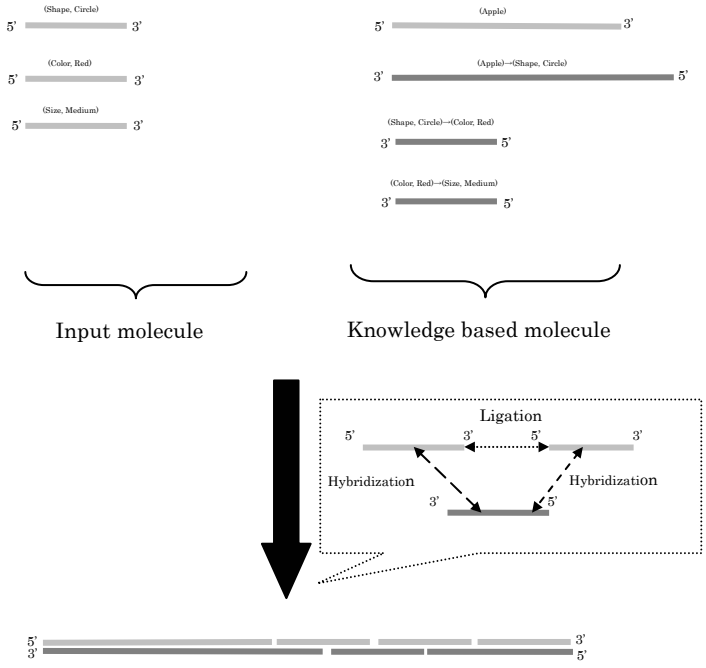


Figure 9: Annealing

3.3 Analysis

The generated DNAs are analyzed into necessary strands which mean double-stranded DNAs representing the target objects. If the necessary strands exist, it comes to that a reasoned reference object is one of the target objects. Generated DNAs are submitted to gel electrophoresis which separates the strands based on size (Figure 10-a). Gel

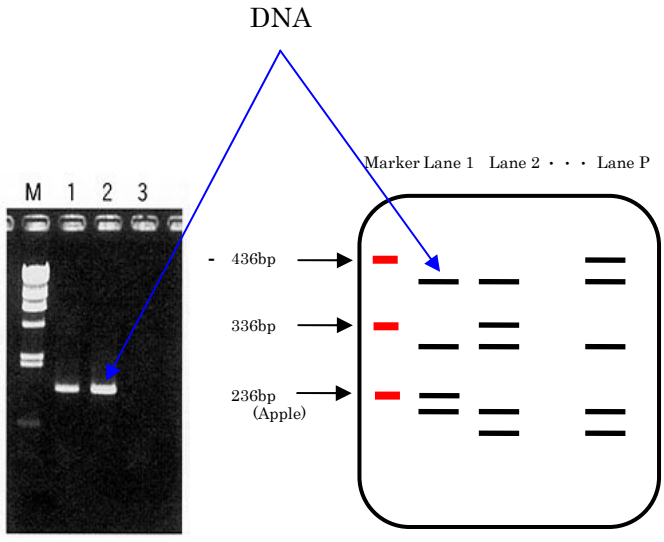
electrophoresis refers to the technique in which molecules are forced across a span of gel, motivated by an electrical current. Activated electrodes at either end of the gel provide the driving force. A molecule's properties determine how rapidly an electric field can move the molecule through a gelatinous medium.

It is possible to measure the size of double stranded DNAs by comparing with DNA markers. All the DNA markers of the same size as necessary double-stranded DNAs are prepared to distinguish generated DNA. Therefore, double-stranded DNAs with the same size as the marker will appear as bands on the gel.

In 1-P tubes, generated DNAs are put into 1-P lanes respectively. After gel electrophoresis is carried out, the result of the analysis is shown in lane 1 to P in Figure 10-b. It is possible to distinguish the objects by their size in respective lanes, because of unique number of the tag vertexes. The size of the necessary double strand is given by N_S. It is denoted as follows,

$$N\_S = S\_D \times N\_A + S\_T \tag{3}$$

where S_D is the size of a DNA piece, N_A ($1 \leq N\_A \leq 7$) is the number of attributes and S_T is the size of the tag. For instance, if a reference object is an apple such as Figure 3, N_D=12, N_A = 3 and S_T = 200, we find out double-stranded DNAs of 236 bp (base pair) exist in lane 1.



10-a: A typical example of gel electrophoresis 10-b: Solution

Figure 10: Gel electrophoresis of reaction

4. DISCUSSION ON DNA COMPUTING APPROACH

In the proposed DNA computing algorithm for the application, there are some chemical processes, PCR, annealing, hybridization, ligation and gel electrophoresis. There are many remarkable papers [1][4][9-12] including this processes. They were primitively used and illustrated as effective methods in DNA computing field. Thus, the proposed algorithm is experimentally feasible.

In this section, the capacity of DNA pieces for a large scale knowledge base is discussed. Moreover the computational complexity is estimated with a simulation result.

4.1 DNA molecules as knowledge bases

In the sequence design, the knowledge (Attribute, Attribute Value) is made a DNA molecule as molecular knowledge, which is the effective in storing a lot of information in knowledge base. One piece, except tag piece, has information of both the attribute and attribute value of the target object. We have to consider an effective way to select the symbol {A, T, C, G} to avoid the error caused by mismatched hybridization which denotes that a sequence is hybridized with non complementary sequence. Recently, the research concerning that has been in progress. Whereas some papers [13-16] report effective methods to resolve the error, in this research we estimate the potential information capacity for proposed model through the produced equation below. We determine the size of a DNA piece representing a tag and/or the knowledge as the unique number and/or twelve nucleotides respectively. These sizes have the limitation of a capacity to store knowledge information. If the number of the target objects and determined attributes increases up to the limitation of the capacity, the knowledge is not fully able to be encoded into twelve nucleotides. Here, we discuss on the size of a piece of DNA molecule with the number of target objects and attributes. In modern computing, it is widely known that the amount of information per unit value is defined as binary codes (0/1). In DNA computing, we determine a unit as symbols. Each of the symbols is selected with the same probability. The number of attribute values is represented by v_{JI} (I =1, 2,..., M ; J =1, 2,…, N) when a target object and an attribute are $O^{(I)}$ and A_J respectively. The amount of information of a DNA piece is given by

$$W(v_{JI}) = -\log_4 \left(\Big/ \sum_{J=1}^{N} \sum_{I=1}^{M} v_{JI} \right) \qquad (2)$$

where $W(v_{JI})$ is a positive integer. In the apple case in Figure 3, let us consider N=3, M=1 and v_{11} = v_{21} = v_{31} = 3. Then $W(v_{JI}) = -\log_4(1/9) = 1.585$. At least, the size of a DNA pieces longer than $W(v_{JI})$ will be required when storing (Attribute, Attribute Value) knowledge information. Moreover we take the errors into account, the size will be still longer. Although for organizations, generally, a DNA sequence is naturally designed to store their genetic information, for molecular knowledge bases DNAs are regarded as a medium of knowledge information. It is possible to store knowledge

information represented by combinations of the each symbol. A double-stranded DNA, one path with a meaning, is formed by bonding some knowledge based molecules with each other. A large scale knowledge base is searched for a specific object (or concept) by the SMD in massive parallelism.

4.2 Evaluation

We might have to evaluate the advantage of the proposed model by using a DNA computer as compared with a silicon-based computer. It commonly says that it is difficult to evaluate a simulation of chemical reaction on the silicon-based computer. DNA-based computers integrate software with hardware and calculate in parallel. If the reaction is simulated on a normal silicon-based computer, it will cause combinatorial explosions depending on the size of a problem. Some of study on artificial intelligence has been studied with regards to avoiding increase in knowledge and computational complexity. From this point, in order to demonstrate the advantage of the proposed model, we estimate computational complexity needed for the solution comparing the DNA-based architecture with simple architecture which means every DNA piece encounters the others in the test tube. It is possible to reason out an object by the combinations between input molecules and knowledge based molecules. In short, the number of the combinations increases with the number of target objects and attributes. Figure 11 shows relations between the attributes and the combinations. X-axis is the number of attributes and y-axis (logarithmic scale) is the number of combinations. On a normal silicon based computer, the number of combinations is carried out by simulating the proposed algorithm with simple silicon-based architecture and with DNA- based architecture separately when there are 3, 100, and 1000 target objects in the molecular knowledge base. With silicon-based architecture, blue, green and red lines are shown in the case of 3, 100 and 1000 target objects respectively. Every three line increases exponentially with the number of attributes. The other hands, with DNA-based architecture, light blue line is shown in all the case of 3, 100, and 1000 target objects. This line increases logarithmically. The number of combinations does not depend on the number of the target objects. We are sure that it was because the proposed DNA computing algorithm did not require complicated mathematical calculation due to DNA self-assembling and DNA-based computers calculate in massive parallelism. The simulation result suggested effective in reducing the computational time under ideal conditions, even if a large scale knowledge base is searched.

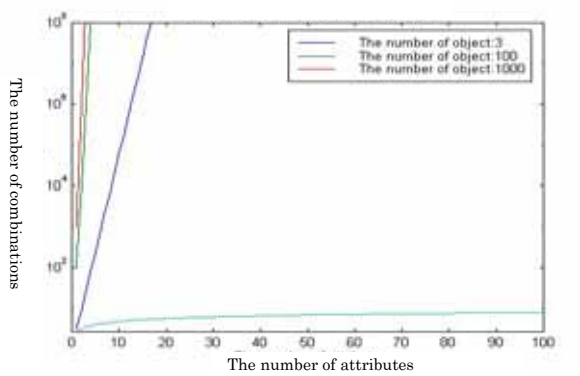


Figure.11 Simulation result

5. CONCLUSION

Self-assembling of molecules mimics the properties of complicated semantic memory. This technique has potential value to analyze a large scale knowledge base thereby in principle. We have presented a new semantic model based on DNA computing. Furthermore, we have discussed on the effectiveness for the proposed model and application. It is demonstrated by following items.

(1) The semantic model is represented by duplex DNAs to be used as a molecular knowledge base.
(2) Reaction and analysis based on self-assembling DNA molecule needed for the solution was outlined with the virtual operations .
(3) We estimated a capacity as a molecular knowledge base to store the knowledge information.
(4) The simulation result demonstrated an advantage of the proposed application by using a DNA-based computer.

Because the proposed model is built from the simple semantic network model, it limits the range of expression to very narrow in comparison with that of expression of man's knowledge. As future works, we are going to build a sophisticated model to express more knowledge. And we hope that the research field of semantic net and artificial intelligence will progress through the proposal of DNA computing approach.

References

[1] L. M. Adleman: Molecular Computation of Solutions to Combinatorial Problems, Science, Vol266, pp.1021-1024, 1994.
[2] J. H. Reif, H. T. LaBean, M. Pirrung, V. S. Rana, B. Guo, C. Kingsford, G. S. Wickham: Experimental Construction of Very Large Scale DNA Databases with Associative Search Capability, The 7th International Workshop on DNA Based Computers, Revised Papers, Lecture Notes in Computer Science 2340, pp.231-247, 2002.
[3] E. Baum: How to Build an Associative Memory Vastly Larger than the Brain Science 268, pp.583-585, 1995.
[4] M. Arita., M. Hagiya and A. Suyama: Joining and Rotating Data with Molecules, IEEE International Conference on Evolutionary Computation, pp.243-248, 1997.
[5] M. R. Quillian and M. Minsky: Semantic Memory, Semantic Information Processing, MIT Press: Cambridge, MA., pp.216-270, 1968.
[6] M. R. Quillian: The Teachable Language Comprehended, A Simulation Program and Theory of Language, Comm. of ACM, Vol. 12, pp.459-476, 1969.
[7] J. H. Rief: Parallel Molecular Computation: Models and Simulations, Proc. of the 7th Annual Symposium on Parallel Algorithms and Architectures, pp.213-223, 1995.
[8] J. Chen, R. Deaton and Y-Z. Wang: A DNA-Based Memory with in Vitro Learning and Associative Recall, The 9th International Workshop on DNA Based Computers, Revised Papers, Lecture Notes in Computer Science 2943, pp. 145-156, 2003.

[9] S. Roweis, E. Winfree, R. Burgoyne, N. N. Chelyapov, M. F. Goodman, P.W. Rothermund, and L. M. Adleman: A Sticker Based Model for DNA Computations, Journal of Computational Biology, Vol. 268, pp.615-629, 1998.
[10] Y. Benenson, T. Paz-Elizur, R. Adar, E. Keinan, Z. Livneh and E. Shapiro: Programmable and Autonomous Computing Machine Made of Biomolecules, Nature, Vol. 414, pp.430-434, 2001.
[11] H. Lim, J. Yun, H. Jang, Y. Chai, S. Yoo, and B. Zhang: Version Space Learning with DNA Molecules, The 8th International Workshop on DNA-Based Computers, Revised Papers, Lecture Notes in Computer Science 2568, pp.143-155, 2003.
[12] M. Yamamoto, N. Matsura, T. Shibata, Y. Kawazoe and A. Ohuchi: Solution of Shortest Path Problems by Concentration Control, The 7th International Workshop on DNA-Based Computers, Revised Papers, Lecture Notes in Computer Science 2340, pp.143-155, 2003.
[13] R. Deaton, C. R. Murphy, M. Garzon, D. R. Franceschetti, and S. E. Stevens Jr.: Good encodings for DNA-based solutions to combinatorial problems, In: DNA Based Computers II, American Mathematical Society, Providence, Vol.44, pp.247–258, 1999.
[14] J. A. Rose, R. J. Deaton, D. R. Franceschetti, M. Garzon, and S. E. Stevens Jr.: A Statistical Mechanical Treatment of Error in The Annealing Biostep of DNA Computation, Proc. of the Genetic and Evolutionary Computation Conference, pp.1829-1834, 1999.
[15] J. SantaLucia, H. T .Allawi, and P. A. Seneviratne: Improved Nearest-Neighbor Parameters for Predicting DNA Duplex Stability, Biochemistry, Vol. 35, No. 11, pp.355-356, 1996.
[16] J. SantaLucia: A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-Neighbor Thermodynamics, Proc. of National Academy of Science U.S.A., Vol. 95, pp.1460-1465, 1998.