Semantic Model for Artificial Intelligence Based on Molecular Computing

Yusei Tsuboi, Zuwairie Ibrahim, and Osamu Ono

Control System 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
{tsuboi, zuwairie, ono}@isc.meiji.ac.jp

Abstract. In this work, a new DNA-based semantic model is proposed and described theoretically. This model, referred to as '*semantic model based on molecular computing'* (SMC) has the structure of a graph formed by the set of all attribute-value pairs contained in the set of represented objects, plus a tag node for each object. Attribute layers composed of attribute values then line up. Each path in the network, from an initial object-representing tag node to a terminal node represents the object named on the tag. Application of the model to a reasoning system was proposed, via virtual DNA operation. On input, object-representing dsDNAs will be formed via parallel self-assembly, from encoded ssDNAs representing (value, attribute)-pairs (nodes), as directed by ssDNA splinting strands representing relations (edges) in the network. The computational complexity of the implementation is estimated via simple simulation, which indicates the advantage of the approach over a simple sequential model.

1 Introduction

Our research group focuses on developing a semantic net (semantic network) [1] via a new computational paradigm. Human information processing often involves comparing concepts. There are various ways of assessing concept similarity, which vary depending on the adopted model of knowledge representation. In featural representations, concepts are represented by sets of features. In Quillian's model of semantic memory, concepts are represented by relationship name via links. Links are labeled by the name of the relationship and are assigned "criteriality tags" that attest to the importance of link. In artificial computer implementations, criteriality tags are numerical values that represent the degree of association between concept pairs (i.e., how often the link is traversed), and the nature of the association. The association is positive if the existence of that link indicates some sort of similarity between the end nodes, and negative otherwise. For example, *superordinate* links (the term used for 'is-a…' relationships) have a positive association, while 'is-not-a…' links have a negative association. Just as there are at least two research communities that deal necessarily with questions of generalization in science, there are at least two bodies of

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3213, pp. 32–39, 2004.

knowledge concerned with representation of the known world as discovered and explained by science. On one hand, knowledge can be fundamentally *procedural and causal*; on the other, knowledge is fundamentally *judgemental* [2]. Naturally, the knowledge representation schemas are quite different; thus, the manner in which the knowledge may be processed to generate new knowledge in each model is also quite different.

Semantic modeling provides a richer data structuring capability for database applications. In particular, research in this area has articulated a number of constructs that provide mechanisms for presenting structurally complex interrelations among data typically arising in commercial applications.

Eric Baum [3] first proposed the idea of using DNA annealing to perform parallel associative search in large databases encoded as sets of DNA strands. This 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's of milliseconds), regardless of the degree of stringency of the recall or the number of simultaneous queries in the input. Arita, *et al.* [4] suggest a method for encoding data and report experimental results for performing concatenation and rotation of DNA. This work also demonstrates the feasibility of join operations in a relational database with molecules. However, this work regarding the database is not based on semantic nets. 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. In this light, we ask: what type of the model is most suitable for implementing such a DNA-based architecture?

In this paper, we propose a new semantic model and its application. The semantic model works on DNA-based architecture, using standard tools from DNA computing. The application is to an Adleman-like [5] scheme which employs primitive motion of DNA strands in the vessel, to effect parallel computation. An important point of his work is the verification of the effectiveness of these approaches via actual experiment.

2 Methodology

In this section, we first provide an overview of the structure of a basic semantic net. Second, we describe how to create a new model, based on DNA molecules. Finally, the proposed model is represented by double-stranded DNAs for purposes of application.

2.1 Structure of Semantic Net

The basic structure of a semantic net is a two-dimensional graph, similar to a network. It is relatively easy for humans to deal with semantic net, because it represents an object (or concept) created from knowledge based on human memories. The semantic net is made of three relations: Object, O; Attribute, A; and Attribute Value, V. In general, this list representation is 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 may be described as a graph with nodes, edges and labels representing their relations. O is reasoned out by the relation between A_i and V_{ji}. Because the semantic net is simply defined with nodes and edges, it is a suitable system to support the search for multiple objects in parallel, and to be used as a knowledge-based system. In general, semantic net size increases with the number of attributes or attribute values.

The other hand, it is imperative to transform complicated graphs into simpler ones. The AND/OR graph enables the reduction of graph size, and facilitates easy understanding. Thus, instead of using the standard existent semantic net described above, in the next section, we instead define a new model, developed to make the most of DNA computing.

2.2 Semantic Model Based on Molecular Computing

First, a tag as a name of an object is set to an initial node in the graph. After we determine the number and kinds of the attribute extracted from the object, both the attribute and attribute value are also set to another node following by the tag node. Second, the relation between nodes and edges is represented using a new defined AND/OR graph. In Fig. 1-a a directed edge in the terminal direction is connected between the nodes in series except for the following case. If there are two nodes which have the same attributes but different attribute values, each of directive edges is connected in parallel as shown in Fig. 1-b. Each edge denotes only connection between the nodes in the directive graph. Finally, labels are attached to the nodes, such as '(Tag)' and '(Attribute, Attribute Value)'.

The nodes denote either a name of the object or both the attribute and attribute value. In short, one path from an initial node to a terminal node means one object named on the tag. We newly define this graph as the *knowledge representation model*. The model represents an object, as reasoned out by the combinations between the nodes connected by the edges. For example, Fig. 2 illustrates this object representation in the context of an apple (named via the tag). An overall graph is then formed by the union of a set of such basic objects, each of which is described in similar, simple fashion. Fig. 3 shows an example of such a network. We name such a graph a *semantic model based on molecular computing* (SMC). An SMC contains all attributes common to every object as well as each attribute value. Attribute layers consist of attribute values, lined up. If an object has no value of a certain attribute, the attribute value is assigned '*no value*'.

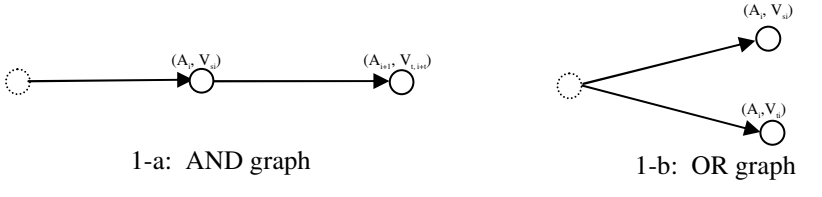

Fig. 1. AND/OR graph connecting nodes in series and in parallel

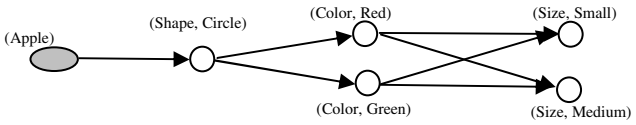

Fig. 2. Simple object model of an apple; The three determined attributes are shape, color, and size

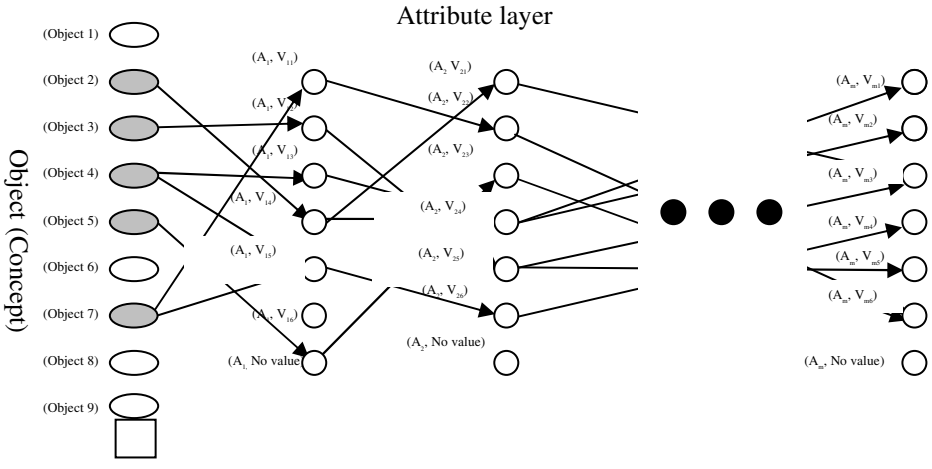

Fig. 3. Semantic Model Based on Molecular Computing (SMC), which collectively models a set of objects, given a total number of attributes, m

2.3 DNA Representation of SMC

Each of the nodes and edges of an SMC may be represented by a DNA strand, as follows: each node (except for tags) is mapped onto a unique, single-stranded (ss) DNA oligonucleotide, in a DNA library of strands. In the DNA library, a row shows attributes, a column shows attribute values and each DNA sequence is designed according to these relations to prevent mishybridization via other unmatching sequences. Every object-naming tag node is represented by a random sequence of unique length (200, 300, 400…) to distinguish the objects. Each edge$_{(Ai,\ Vis)\rightarrow(\bullet Ai+1,\ Vi+1,\ s)}$ from node$_{(Ai,\ Vis)}$ to node$_{(Ai+1,\ Vi+1,\ t)}$ is designed to be Watson-Crick complementary to the node sequences derived from the 3' 10-mer of node$_{(Ai,\ Vis)}$ and the 5' 10-mer of node$_{(Ai+1,\ Vi+1,\ t)}$. Except for initial and terminal edge strands of each graph path, each is a ssDNA oligonucleotide of length 20. These two ssDNAs are respectively represented by the size which suits the end of the DNA pieces of the initial or the terminal node exactly.

In this way, the SMC is represented by double-stranded (ds) DNAs. Fig. 4 shows one of the paths shown for the apple model in Fig. 3, as represented by a dsDNA ((Apple)→(Shape, Circle)→ (Color, Red)→ (Size, Medium)).

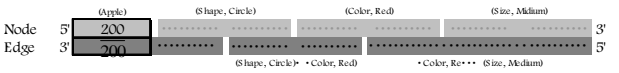

Fig. 4. One of the double- stranded DNAs represented by the graph (Apple) in Fig. 2

3 Application

The following demonstrates the application of the semantic model to a reasoning system. The system is implemented by chemical operations with DNA molecules.

3.1 Reasoning System

This reasoning system consists of: (a) Input, (b) Knowledge base, (c) Reasoning engine, and (d) Output.

a) Input
In the input, the attribute values are extracted from an *input objec*t separately according to previously determined attributes. Using the attributes and attribute values, a ssDNA is synthesized as an *input molecule*.
b) Knowledge base
In the knowledge base, a ssDNA representation of each edge and tag in the network is synthesized as a *knowledge based molecule*.
c) Reasoning engine
The reasoning engine denotes the biochemical reactions which occur under experimental conditions, given a particular set of input molecules, and the complete set of knowledge based molecules.
d) Output
Output refers to the dsDNA products, as determined via length.

3.2 Implementation

In this work, the system is implemented by virtual chemical operations. For reasonable operation, each of the knowledge based and the input molecules must first be amplified sufficiently. Knowledge based molecules are inserted into a test tube as a *molecular knowledge based memory*. Input molecules are then put into the test tubes. It is assumed that each ssDNA will spontaneously anneal to complementary sequences under defined reaction conditions in the test tube. The ssDNA sequences representing input molecules and knowledge based molecules are then mixed in the presence of DNA ligase, which will form a covalent bond between each template-directed pair of adjacent DNAs, directed by a pair of complementary single-stranded overhangs. Thus, each set of sequences is ligated to form a dsDNA which represents a path between an initial node and a terminal node in the model. As a result, all possible dsDNAs representing the paths are generated at random.

The generated dsDNA set must then be analyzed to determine the specific set of represented objects, as produced by the reaction. Generated dsDNAs are subjected to

gel electrophoresis, which separates the strands based on length, which then appear as a discrete bands on the gel in a lane.

The length of each generated dsDNA, denoted as N_S, is given by the simple relation:

$$N\_S = L\_D \times N\_A + L\_T, \tag{2}$$

where L_D is the length of ligated dsDNA, except for the tag sequence, N_A is the number of attributes, and L_T is the length of the tag. For instance, if a reference object is an apple such as L_D=20, N_A = 3 and L_T = 200, we find out double-stranded DNAs of 260 bp (base pair) exist in the lane.

4 Discussion

The model and implementation presented in this paper relies on chemical processes such as annealing and gel electrophoresis. In actual practice, an effective way to select sequence to avoid mismatched, error hybridization will have to be devised. Recently, substantial progress has been reported on this issue [6]–[8]. We expect that this issue will be resolved satisfactorily in the near future.

The proposed model is applied to knowledge based memory via DNA molecules, which is in some sense similar to human memory, due to the inherent massive parallelism. This performance is not realized in artificial, sequential models of computing. Although simulations will be interesting, the inherent advantages provided by the design will therefore be evident only when using real DNA molecules.

We might have to evaluate the advantage of the proposed model by using a DNA computer as compared with a silicon-based computer. It is commonly said 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. A direct attempt to simulate the implemented reaction on a normal silicon-based computer will be compromised by the potential for a combinatorial explosion in the number of distinct path molecules. Some studies on artificial intelligence have been performed with regards to avoiding such increases in knowledge and computational complexity. For this reason, in order to demonstrate the advantage of the proposed model over a simple, sequential model, we estimate the computational complexity required for solution, assuming that every ssDNA encounters all others in the test tube. It is possible to reason out an object by the combinations between input molecules and knowledge based molecules. Therefore, it is reasonable to expect the number of combinations to increase with the number of objects and attributes. Fig. 5 shows relations between the attributes and the combinations. The number of combinations is estimated for the simple, sequential architecture and a DNA-based architecture separately when there are 3, 100, and 1000 target objects in the molecular knowledge based memory. With a simple architecture, blue, green and red lines are shown, which correspond to the case of 3, 100 and 1000 objects respectively. Each of these three lines (Only three of 4 lines are labeled in the figure; This should be corrected…) increases exponentially with the number of attributes. In contrast, a single light blue line indicates the operation number required for a DNA-based architecture

for each of the case of 3, 100, and 1000 objects. This line also increases exponentially with attribute number. However, the number of combinations does not depend on the number of target objects, since the proposed application requires only DNA self-assembly which proceeds for all objects in parallel. This simulation result suggests that the proposed implementation will be effective in reducing the computational time, under ideal conditions.

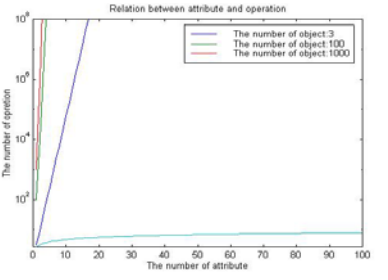

Fig. 5. Estimation of the computational complexity, with increasing number of attributes and objects in the knowledge based memory

5 Conclusion

In this work, a semantic model has been presented for knowledge representation with DNA molecules, as follows:

(1) A newly-defined semantic model, in which objects are represented by dsDNAs;
(2) For the stated application, reaction proceeds via DNA self-assembly. This process was outlined, and analyzed via simulation, from a theoretical point of view.
(3) We estimated the computation complexity of the DNA-based architecture, and compared with that of a simple, sequential architecture;

Since the inception of DNA-based computing, a primary concern has been the development of applications in the field of engineering and artificial intelligence. The proposed model suggests that DNA-based computing should be applicable to the artificial intelligence field. It seems likely that this approach will be utilized as a natural application for problems involving pattern matching, deduction, and retrieval in the future.

Acknowledgement

The authors would like to thank J. A. Rose of the University of Tokyo for helpful comments that led to many improvements in this paper.

References

1. Quillan, M.R.: Semantic Memory. In Semantic Inform. Processing, M. Minsky, Ed. Cambridge, MA: MIT Press (1968)
2. Blanning, R. F.: Management Applications of Expert Systems. Information and Management, Vol.7 (1984) 311-316
3. Baum, E. B.: How to Build an Associative Memory Vastly Larger than the Brain. Science 268 (1995) 583-585
4. Arita, M., Hagiya, M., and Suyama, A.: Joining and Rotating Data with Molecules. IEEE International Conference on Evolutionary Computation, pp.243-248, (1997)
5. Adleman, L. M.: Molecular Computation of Solutions to Combinatorial Problems. Science, Vol.266 (1994) 1021-102 4
6. Deaton, R., Murphy, C. R, Garzon, M., D. R. Franceschetti and S. E. Stevens, Jr.: Good Encodings for DNA-based Solutions to Combinatorial Problems. DNA Based Computers II DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol.44 (1999) 247-258
7. Rose, J. A., Deaton, R., Franceschetti, D., Garzon, M., and Stevens, S. E. Jr.: A Statistical Mechanical Treatment of Error in the Annealing Biostep of DNA Computation., Proc. GECCO'99, pp.1829-1834 (1999)
8. SantaLucia, J., Allawi, H., Seneviratne, P.: Improved Nearest-Neighbor Parameters for Predicting DNA Duplex Stability. Biochemistry, Vol.35, No.11 (1996) 355-356
9. Garzon, M., Bobba, K., Neel, A.: Efficiency and Reliability of Semantic Retrieval in DNA-based Memories, Lecture Notes in Computer Science, Springer-Verlag Heidelberg, pp. 379-389 (2003)