DNA Computing and Its Application on NP-Completeness Problem

Rofilde Hasudungan[1], Rohani Abu Bakar[2], Rozlina Mohamed[3]
[123] Faculty Computer System and Software Engineering, Universiti Malaysia Pahang, Gambang 26300, Malaysia
[1]mcs11001@stdmail.ump.edu.my, [2]rohani@ump.edu.my, [3]rozlina@ump.edu.my

Abstract. DNA computing is emerging research area that attracts many researchers in multidiscipline area range from biology, physics, mathematics, and computer science. Nowadays, many researchers already solved problem using this technique, mostly classified as NP (Nondeterministic Polynomial) problem since the inheritance of DNA computing in processing simultaneously and it huge memory capacity. Despite very promising, DNA computing faces several obstacles such as exponential solution explosion, representing weight etc. This paper aim is to give a review on current achievement on DNA computing upon NP-completeness problem.

Keywords: DNA computing; NP problem; DNA strands; Data representation on DNA strands

1. Introduction

Since introduced by Adleman [1] through wet experiment in solving one of instance HPP, DNA computing has attract many researcher in multidiscipline area, Adleman pointed out the advantages doing computation in DNA regarding its capabilities to process simultaneously. In his experiment, Adleman utilize DNA Computing to solve traveling salesman problem. To solve this, Adleman represent cities into DNA strands, conduct bio-chemical technique in wet experiment to form solutions, and sort solutions to choose the best one.

The main idea behind DNA based computer is to utilize DNA as media to store information and use biological technique as computational process. As media to store information, DNA computing is very density molecule so as it is possible to store huge of data in DNA strand. There are 3 advantages using DNA based computer to conduct computation there are

1) Speed. The speed of any computer, biological or not, is determined by two factors: (i) how many parallel process it has; (ii) how many steps each can perform per unit time. The exciting point about biology is that the first of these factors can be very large; recall that a small amount of water contains about 10^{24} molecules.
2) Information stored in DNA strand. DNA molecule is very dense. In conventional computer, data represent by binary 0, 1 compare with 4 letters (A, G, C and T) in DNA based computer. Now days, it's possible to store 5.5 petabits of data or

around 700 terabytes in single gram of DNA, smashing the previous DNA data density record by thousand times [4].

3) Energy efficiency [2].
DNA based computer is very efficient in term of energy used for computation, it just require one joule to perform $2\ x\ 10^{19}$ligation.

Based on the advantages using DNA based computer to solve NP-completeness problem many researcher applying several of problem such as clustering [5,6,7,8], scheduling [8,7,9], graph theory [1,10] and soon.

In this paper we extend a review current DNA computing technology, encoding weighted in DNA strands and achievement mainly on solving NP-completeness problem. The paper organized as follows. In section 2, we will represent brief introduction DNA molecule. Section 3 describes biotechnology technique used in DNA computing. Section 4 addresses current achievement in NP-Complete problem or combinatorial problem. Representation numerical value in DNA computing will address in section 5. Conclusion and feature work will give in section 6.

2. Molecular Computing

DNA is a molecule that carries information from generation to generation. DNA is a crucial molecule in living cells (*in vivo*) and it has fascinating structure which is support two important functions of DNA: coding for the production of proteins, and self-replication so that an exact copy is passed to the offspring cells. Structurally, DNA molecule is a polymer constructed from series of monomers called nucleotides. Nucleotides consist of three elements there are: *a group of sugar, a group of phosphate* and *bases*. Nucleotides only differ in their bases. There 4 bases attached in DNA molecule: Guanine, Cytosine, Adenine and Thymine, abbreviated as G, C, A and T respectively. Each strand, according to chemical convention, has a 5' and a 3' end; thus, any single strand has a natural orientation. This orientation is due to the fact that one end of the single strand has a free (i.e., unattached to another nucleotide) 5' phosphate group, and the other end has a free 3' deoxyribose hydroxyl group.

Naturally, nucleotides can be interacting to form bonds. There are two bonds: covalent bonds and hydrogen bonds. The covalent bonds is when two strands DNA interact and join together phosphate disaster bond to form new DNA strands, this process called ligation. This ligation happen involved enzyme called ligase, where ligase enzyme will act as "glue" to seal bond between strands. Meanwhile, hydrogen bond occur when DNA strand interact with its complement. This process will form double helix DNA and the bonding will be following Watson-Crick Complementarity and the process called hybridization. Watson-Crick Complementarity rules are: A will pair with T with 2 hydrogen bonds meanwhile C will pair with G with 3 hydrogen bonds, and no other possibility for bonding. Besides that, a strand will only anneal to its complement if they have any opposite polarities. One strand of double helix extends from 5' to 3' and the other from 3' to 5'.

Synthesis
Synthesis is process designing and translation information into DNA sequence form. In DNA computing this process is involved designing encoding schemes and synthesized data based on that design.

Denaturizing, annealing and ligation
When DNA double helix is melting in temperature 85^0-95^0 C (depend on the containing GC pairs), it will be separated into two strands, this process called *denaturation*. Otherwise, *annealing* (also known as hybridization) is reverse of *denaturation,* is process that fusing two single stranded molecule by complementary base pairing. The process is cooling down the solution; separated strands fuse by the hydrogen bonds. Meanwhile, Ligation is a process that joins two DNA molecule ends from the same or different molecules. This process involves creating phosphodiester bond between the 3' OH of one nucleotide and 5' phosphate of other nucleotides. The reaction is catalyzed by DNA ligase enzyme. Ligase enzyme is used as "glue" to seal the covalent bonds between the adjacent fragments.

Gel Electrophoresis
Gel electrophoresis is technique used for sorting DNA strands by its size [11]. Since DNA molecule carries a negative charge to sort. Since DNA molecules carry a negative charge, when placed in an electric field they tend to migrate toward the positive pole. The rate of migration of a molecule in an aqueous solution depends on its shape and electric charge. Since DNA molecules have the same charge per unit length, they all migrate at the same speed in an aqueous solution. Smaller molecules therefore migrate faster through the gel, thus sorting them according to size.

Primer Extension and PCR
PCR is copying machine for DNA at the same time can be used for DNA detection. PCR capable to copy a million or even billion of similar molecules based on single specific molecule. Each cycle of the reaction doubles the quantity of each strand.

3. Achievement in Solving NP-Completeness Problem

After Adleman's experiment there are many researcher applying DNA computing in various area, mainly applying DNA to solve NP-Complete problem. Lipton showed the advantages huge parallelism inherent in DNA based computing through solving "satisfaction" problem (SAT) [12].Consider the following formulas

$$F = (x \vee y) \wedge (\bar{x} \vee \bar{y}).$$

The variable *x* and *y* are Boolean; they are allowed to range only over two values 0, 1. The SAT problem is to find Boolean values for *x* and *y* that make the formula *F* true.

Ouyang *et al*. [13] represent DNA computing capabilities to solve maximal clique problem. A clique is defined as a set of vertices in which every vertex is connected to every other vertex by an edge. The maximal clique problem is to find the largest clique and the vertices on it where given a network contain *N* vertices and *M* edges.

Bakar and Watada use DNA computation to solve clustering problem based on mutual order [7]. The proposed method used Euclidean to group the pattern, given number of point, counting all of distance between point using Euclidean function, sort all value and make order into value. Represent all point and distance order into DNA strand. Through biological technique search all possible group of point and search for the best of them. Meanwhile Bakar and Watada solve clustering problem in DNA computing based on k-means and fuzzy C-means, Zhang and Liu work at a CLIQUE algorithm [8]. Zhang and Liu proposed closed-circle method in DNA computing. In proposed method, the process clustering becomes a parallel bio—chemical reaction and the DNA sequences representing marked cells can be combined to form closed-circle DNA sequences. CLIQUE algorithm used to cluster high-dimensional data, this algorithm based on density and grid. The CLIQUE algorithm first finds one-dimensional dense grids, then two dimensional dense rectangles and so on, until all dense hyper rectangles of dimensionality k are found. In proposed method, Zhang applying proposed method in two dimensional data.

Bakar and Watada applying DNA based computer to solve determination of logistic problem [5,6], this problem could be class as NP Hard. In this study Bakar proposed DNA encoding schemes to represent data of retailer cities and distribution center and used bio-chemical technique as computation tools. In order to represent weight or distance between retailer city and distribution center Bakar employed proportional length method [14] and melting temperature technique [15].

Rearrangement scheduling robotic cells at flexible manufacturing has proposed by Bakar and Watada [9]. Given 6 machine that provide part of product, three assemble table that assembly three different product. The robot will move in inline production line in order to produce product. In this study, the optimal movement of robot in order to produce product in minimum time is investigated.

Solving maximal clique problem by proposing model based on circular DNA length growth. This technique used circular single stranded DNA (c-ssDNA) and it is implemented to the computation on nano-scale; the target DNA (circular single stranded DNA) molecules will growth in length just when they satisfy the conditions. Based on experiment, the maximal clique problem can be solved using $O(m+n)$ operational time complexity. Proposed method has advantages: (1) CDLG used to select feasible solutions; (2) less computational complexity. Jing and Xu believe that proposed algorithm can be implemented on another NP-Complete problem such as SAT, and maximal independent set problem. Meanwhile, Zoraida [15] solve Chine postman problem (CPP) by utilize the thermodynamic properties of DNA strand as parameter to encoding numerical value for distance. CPP is problem to find the shortest route in a network that use every arc (directed edge) and gets back to where they started; its alike salesman traveling problem.

Along with increasing number of data used in computation effecting exponential solution space explosion in DNA computing and become big obstacle in DNA computing. To overcome this issue Xu *et al*. proposed new algorithm that capable to decrease volume exponential explosion solution in DNA computing for solving vertex cover problem. The algorithm based on sticker-based model for generating solutions and Adleman-Lipton model for biological operation. The proposed algorithm shown that the used of strands for solution have been decrease from $O(2^n)$ long DNA strand to $O\sqrt{3^n}$. 3D DNA self-assembly employed by Zhang et al. for maximal clique

problem. Zhang use complementary graph to design seed configuration and 3D titles, and use DNA self-assembly to decide which sets of vertices are clique. The proposed method has capability to decrease the steps of manual intervention and reduce error rate as well. Xu et al proposed computing model using DNA computing to solve 3-colorable graph with 61 vertices. The model consists of three main steps: (1) Subgraph division, (2) Deleting false solution of each Subgraph, and (3) finding the true solutions. This model has capability to delete more than 99% false solutions as so the searching capability of this model could be up to $O(3^{59})$ where this is the largest among electronic and non-electronic computers. Meanwhile, Xu et al. proposed unenumerative DNA computing model for graph vertex coloring problem. This technique objective is to avoid solution space exponential explosion caused by the enumeration of the candidate solution. In this technique, Xu et al. employed two techniques: (1) ordering the vertex sequence for given graph (2) reducing number of encoding representing color according to the construture of given graph. Experiment of proposed method showed that is with 12 vertices in graph without triangles is solved and its initial solution space included only 283 DNA strands, which is 0.0.532 of 3^{12}.

Molecular beacon-based DNA computing model was proposed by Zhixiang et al. for solving maximal independent set problem. In proposed algorithm, molecular beacon used as probe, to separation and detection all of the independent sets and by observing the fluorescence to determine whether or not the solution is existence. This model also does not require the restriction enzyme digestion, gel electrophoresis etc. these steps may avoid possible computational errors and data loss. Meanwhile, Wang et al. by utilizes plasmid and a special separation device. Plasmid is used as screen out the efficient solutions and deletes the non-solution. Meanwhile, a special separation devise is used to pick-up true solution since it cans pick-up solution efficient and fast. According experimental measures and feasibility of biomaterial and bio-methods the proposed method is efficient and feasible.

Razzazi and Roayaei using sticker model to present three DNA algorithms for solving three different NP-complete graph-based problems for the first time: domatic partition, kernel and induced path. The sticker model has four basic operations: merge, separate, set, and clear. Computer simulation is conducting to prove the correctness of propose method.

4. Represent Numerical Value in DNA Computing

The numerical data representation method in DNA strands is the one of the important issue to extend the field of DNA computing. Though some researchers tried to represent the numerical data using DNA, the results are not satisfactory yet. A Mistake in encoding numerical value will result wrong solutions and error in process. Many researcher have been proposed how to represent numerical values in DNA strands, most of them utilize characteristic of DNA strand itself such as number of hydrogen bond, temperature, GC content, length of strand and so forth.

4. 1. Proportional length method

When Adleman using DNA computing for solving instance of TSP, he represent weight directly based on its value. Hartamanis and Amos *et al.* pointed out the inefficient method, its required huge amount DNA to form initial set of routes. Hartamanis also count, to solve TSP with 200-city given required a mass of DNA greater that the earth. To overcome those problems Narayanan proposed algorithm that called proportional length method. In this method the numerical value is represented in proportional way. For example, if there are 4 routes R = {2, 4, 6, 8}, if R=2 is represented by strand of length 2 than R=4 will be represented by strand of length 4.

4. 2. Fixed code-length method

While Narayanan and Zorbalas [14] proposed proportional length based to encode numerical value, Shin *et al.* [17] proposed fixed code-length. Even proportional length based feasible to represent numerical value and easy to read, but in some case it become inefficient. If constant factor k=3 and there has numerical value {1, 2, 100}, represent strands in that value become inefficient. To overcome this drawback, Shin utilizes G/C pairs in strands to represent numerical value while length of strands is represented as fixed length. By using genetic algorithm the amount of G/C content in edge sequences is optimized, so the edge with smaller weights have more G/C contents and have higher probability of being contained in the final solution. To evaluate the proposed method simulation is done to solve traveling salesman problem. Shin believes proposed method has promising to be implemented in large-scale problem.

4. 3. Concentration method

Yamamoto *et al*. proposed concentration control method to encoding numerical value for DNA computing [18]. The proposed method represents weight by relative concentration of each nucleotide and used it as input and output in this model. By using this technique, it is possible to do local search instead of exhaustive search since hopeless candidate solutions tend to small.

4. 4. Temperature gradient method

Improve the previous method [17], Lee *et al.* proposed temperature gradient to encode numerical value in sequences [15]. In this proposed method, the melting temperature (Tm) becomes the key how the numerical value to be represented, since it is determined by more complex function consists of various factors including GC content. The method used melting temperature to control hybridization process. In this method the weight sequence with the smaller weights have a lower *Tm*.

4. 5. Hybrid Concentration Control Direct-proportional length-based

Ibrahim *et al*. [17] proposed concentration control directed-proportional length-based method to encode numerical value into DNA strands. This method is improvement of proportional length method that has been proposed by Narayanan and Zorbalas [14]. This method based on combination of two characteristics: length and concentration, for encoding and at the same time, effectively control the degree of hybridization of DNA. Length is used to encode the cost of each path in proportional way; meanwhile, the hybridization control by concentration is done by varying the amount of oligo, as input of computation, before the computations begin.

The protocol used in this proposed method similar as direct proportional based method, but the difference is that the amount of poured DNA representing the edges varies closely to the weight of edges. As a result, concentration of input during initial pool generation will be different as well, and does influence the degree of hybridization.

4. 6. Temperature control method

Li *et al.* proposed a melting temperature control encoding method. The method uses fixed-length DNA strands and represents weights by melting temperatures (T_m) of the given DNA strands. The basic idea of this proposed method is to design the sequences so that the DNA strands for higher-weight value have higher melting temperature than those for lower-weight values. The weight sequences describe the proportion of edge weights, so this coding scheme can express the real value weights, and therefore a more economical path has a lower G/C contents. Each vertex sequence is designed to have the same melting temperature, because the vertex sequences should contribute equally to the thermal stability of path.

4. 7. Incomplete Molecule Commixed Encoding (IMCE)

Wang *et al.* [19] proposed Incomplete Molecule Commixed Encoding (IMCE) to encode numerical value into DNA strand; this encoding schemes encoding vertex, weight and edge in different ways. In IMCE model, incomplete model is used similar with domino where used to encode edge. The form of in complete molecule consisted with three parts: half single stranded of vertex *i*, double-strand of weight *ij* and single-strand of vertex *j*. Meanwhile, vertex encoding represent with single stranded-encoding and weight encoding is based on distribution ratio.

4. 8. Order number of weight and relative length graph

Xikui and Yan proposed method for encoding weight based on number of weight and relative length graph. Proposed method encode vertex consisting of position and weight. Since vertex is represented in fix length DNA sequences, represent weight in vertex by varying number of A/T pairs and G/C pairs. Generally, DNA length and G/C contents influence the ligation between sequences and more G/C pair in the

sequence has, the more probable they get hybridized. To optimize the edge weight sequences used fitness function where edges with small weight have more G/C contents.

5. Conclusion and future work

Based on literature, there are many problem already successfully solved using DNA computing. Massive parallelism and memory capacity become this technique possible and very promising. Even very promising and suitable to solve complex problem but there has drawback needed to be solved (1) Representing weight is one of main problem in DNA computing is very important. Though some researchers tried to represent the numerical data using DNA, the results are not satisfactory yet. (2) In literature there has many problems solved in DNA computing using several model and technique in order to increase capability and feasibility DNA computing, but most of them just applied in specific problem become it limited. Generalize the proposed method will safe more time and cost.

Acknowledgments

This research is funded by ERGS Grant: An Exploration on the Generic DNA Sequence Design Scheme to Solve Engineering Application Problem (RDU110603).

6. References

1. Adleman, L.: Molecular computational of solutions to combinatorial problem., 1021-1024 (1994)
2. Church, G., Gao, Y., Kosuri, S.: Next-Generation Digital Information Storage in DNA., 1628 (2012)
3. Ito, Y., Fukusaki, E.: DNA as a 'Nanomaterial'. Journal of Molecular Catalysis B: Enzymatic 28 , 155-166 (2004)
4. Bakar, R., Watada, J.: A DNA computing approach to cluster-based logistic design. (2007)
5. Bakar, R., Watada, J.: A biologically inspired computing approach to solve cluster-based determination of logistic problem. International Journal of Biomedical Soft Computing and Human Science 13, No. 2, 59-66 (2008)
6. Bakar, R., Watada, J., Pedrycz, W.: DNA approach to solve clustering problem based on mutual order., 1-12 (2008)
7. Zhang, H., Liu, X.: A CLIQUE algorithm using DNA computing techniques based on closed-circle DNA sequences., 73-82 (2011)
8. Bakar, R., Watada, J.: A bio-soft computing approach to re-arrange a flexible manufacturing robot., 308-315 (2007)

9. Xue, J., Liu, X.: Applying DNA computation to clustering in graph. In : International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), pp.986-989 (2011)
10. Ouyang, Q., Kaplan, P., Liu, S., Libchaber, A.: DNA solution of the maximal clique problem. (1997)
11. Amos, M.: Theoritical and experimental DNA Computation. Springer (2004)
12. Paun, G., Rozenberg, G., Salooma, A.: DNA computing: new computing paragims. Springer, New York (1998)
13. Lipton, R.: Using DNA to solve NP-Complete problems., 542-545 (1995)
14. Narayanan, A., Zorbalas, S.: DNA algorithms for computing shortest paths. Genetic Programming , 718-723 (1998)
15. Lee, J., Shin, S., Augh, S., Park, T., Zhang, B.: Temperature gradient-based DNA computing for graph problem with weighted edges. DNA 8, LNCS 2568, 73-84 (2003)
16. Zoraida, B.: DNA algorithm employing temperature gradient for chinese postman problem. In : International Conference on Process Automation, Control and Computing (PACC), pp.1-4 (2011)
17. Shin, S., Zhang, B., Jun, S.: Solving traveling salesman problem using molecular programming. In : Proceedings of Congress on Evolutionary Computation 1999, pp.994-1000 (1999)
18. Yamamoto, M., Matsuura, N., Shiba, T., Kawazoe, Y., Ohuchi, A.: Solutions of shorthest path problems by concentration control. DNA Computing: 7th International Workshop on DNA-Based Computers, 203-212 (2002)
19. Ibrahim, Z., Tsuboi, Y., Ono, O.: Direct-proportional length-based DNA computing for shortest path problem. International Journal of Computer Science and Applications I, No. 1, 46-60 (2004)
20. Han, A., Zhu, D.: A new DNA encoding method for traveling salesman problem. In : ICIC, pp.328-335 (2006)
21. Wang, Q., Pei, Z., Hou, X., Sun, Q., Zheng, H.: DNA algorithm based on incomplete molecule commixed encoding for the shorthest path problem. In : International Conference on Information Science and Engineering (ICISE), pp.3641-3644 (2009)
22. Hartamanis, J.: On the weight of computations. Bulletin of the European Assosiation for Theoritical Computer Science 55, 136-138 (1995)
23. Lee, J., Shin, S., Park, T., Zhang, B.: Solving traveling salesman problems with DNA molecules encoding numerical values. BioSystem 78, 39-47 (2004)
24. Li, Y.: Traveling salesman problem based on DNA computing. In : Third International Conference on Natural Computation (ICNC) (2007)
25. Xikui, L., Yan, L.: DNA computing for Traveling Salesman problem. In : 3rd International Conference on Bioinformatics and Biomedical Engineering, pp.1-4 (2009)