Searching DNA Decision Trees Using Quantum Dots Detection

Jan J. Mulawka

Warsaw University of Technology, Institute of Electronic Systems, Warsaw, Poland,
e-mail: jml@ise.pw.edu.pl

Abstract. Concept of self-assembly of DNA molecules may be utilized to implement different algorithms of computing. In particular, this methodology is useful in artificial intelligence because it operates on symbols. As has been shown by the author [7] DNA computing is suitable for searching the decision tree which is encoded by DNA molecules. Thus, the algorithms of artificial intelligence can be performed by this technique. In the contribution we demonstrate that by proper encoding and manipulating with DNA molecules it is possible to implement reasoning procedure by searching the decision tree. The method uses standard genetic engineering operations like hybridization, ligation, and purification. Quantum dot technique is applied to speed up detection of the final output eliminating the slow operation of electrophoresis. The approach presented is very simple and fast because the technique used allows an enormous number of molecules to be labeled, reduces instrument tie-up and improves analysis throughout the process.

1 Introduction

Recently, new initiative of research in nanotechnology has emerged. Due to limitations of semiconductor and microprocessor technology new paradigms of computing are suggested. One solution to this problem is to apply a methodology developed in molecular genetics [1] and to use molecules of DNA for manipulating and storing the data. Such approach is known as molecular computing [9-11].

DNA molecules are linear polymers called oligonucleotides or strands. They are composed of single or double strands. A single-strand has a phospho-sugar backbone and four bases denoted by the symbols A, T, C, and G. Thus, DNA strands are equivalent to strings composed of letters which belong to four-element alphabet. In molecular computing self-assembly is an essential operation on strands. As a result a double-stranded molecule is formed by two single-stranded molecules due to the hybridization reaction, because A is complementary with T and C is complementary with G [1]. The DNA strands are found to be adequate for symbolic computation, in a way similar to that of symbol-processing languages for formalized representation of knowledge.

The area of artificial intelligence and especially expert systems [2] seem to be well suited for this purpose. Such systems simulate the problem solving processes as of a human expert. By reasoning or knowledge processing we mean deriving a conclusion, or conclusions from a knowledge base using some strategies and procedures. Generally, the inference process involves a search procedure. In some cases, the search moves in a forward direction - from premises (or facts) to conclusions. In others, the search moves backward – from a hypothesized conclusion to the premises necessary to infer that conclusion. Since operations of molecular computing are performed in genetic engineering laboratory, inferencing based on DNA molecules is some kind of experimental artificial intelligence.

The first results on molecular computing were reported by Adleman [12]. He demonstrated that NP-complete combinatorial and graph problems, which are difficult for traditional computers, may be solved using this method. However, DNA computations require algorithms that are quite different from those used in conventional computers [13-27].

As was shown by Mulawka and Węgleński the inference process may be performed at the molecular level [3-7]. The encoding of the knowledge base may be done by adequate DNA strands which can be handled in a standard DNA laboratory. Implementation of the inference engine applied in [7] is based on a decision tree algorithm. The disadvantage of this approach is the necessity of applying complicated operations of genetic engineering such as autoradiography as well as complicated identification of decision paths by their lengths using electrophoresis. However paradigm of reasoning may be significantly simplified by applying recent achievements of nanobiotechnology.

In this contribution we present the DNA implementation of the inference engine based on decision tree with identification of the final result by using quantum dots [28]. The idea is explained on a knowledge base with a decision tree of 12 nodes and 11 branches which is equivalent to seven rules and five initial facts. The approach uses standard genetic DNA engineering technique applying the following operations: hybridization, ligation, and identification of molecules by quantum dots. It is described how the algorithm of reasoning is encoded and performed.

2 Quantum dots

Quantum dots are colloidal semiconductor nanocrystals [28-31] which exhibit unique optical and electronic properties due to the confinement of electronic excitations. In addition, they combine the advantages of high photobleaching threshold with good chemical stability. Essential parts of the quantum dot are created by core and shell. The core is composed of materials such as cadium selenide, cadium telluride, or indium arsenide. The core material is selected to coarsely control the emission wavelength. The size of the spherical core determines the optical properties of the quantum dot and is used to fine tune the exact wavelength. Fluorescent emission can be tuned across the visible spectrum by simply changing the particle radius.

The core nanocrystals are overcoated with a shell built of another inorganic material. The shell serves to protect the core, amplify the optical properties as well as insulate the core from environmental effects. To link the shell to nucleic acid a third layer composed of an organic surface coating is used. The quantum dot shell retains its biochemical properties by coupling to it DNA oligonucleotide.

Quantum dots are used to produce fluorescent tags for DNA oligonucleotides. These tags may be applied effectively to detect results of bioassays as provided in Fig. 1. By using adequate optical equipment this technique enables the labeling and detection of an enormous number of DNA molecules.

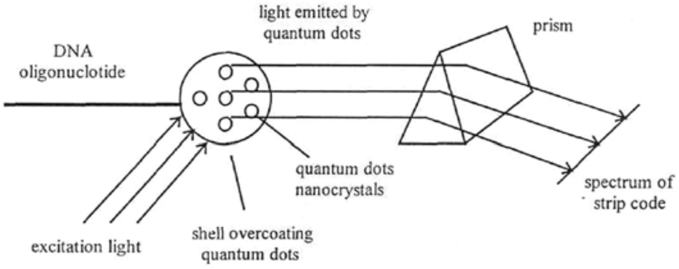

Figure 1. Application of quantum dots to detect DNA oligonucleotides.

Thus, quantum dots have the potential to overcome many problems encountered by organic small molecules in certain fluorescent tagging applications. In particular, fluorescent techniques can be performed much faster than conventional enzymatic and chemiluminescent techniques, reducing instrument tie-up and improving assay throughput. Therefore it is interesting to apply this new way of biochemical molecules labeling in molecular computing.

3 Decision trees

Production rules are popular for knowledge representation within expert systems [2]. The rules may be graphically represented by so called inference networks. These networks comprise assertions and intermediate conclusions which may be combined by logical connectives. In our method of knowledge representation we apply a decision tree which is generally more simple kind of the inference network. The structure of such a tree forms a hierarchic graph as depicted in Fig. 2.

Generally, the decision tree may be considered as an acyclic directed graph comprising some number of nodes and branches. In such a graph we can distinguish three kinds of nodes: initial node - so called root, intermediate nodes and final nodes known as leafs. The node at the highest level in the tree is the root while at the lowest level - the leaf. The branches in this graph may emanate from the root and from intermediate nodes, while a number of branches may be arbitrary. Each node in the decision tree represents either a question about the value of an attribute, or a conclusion. Each branch that emanates from a node pertaining to a question will represent one of possible values of the associated attribute.

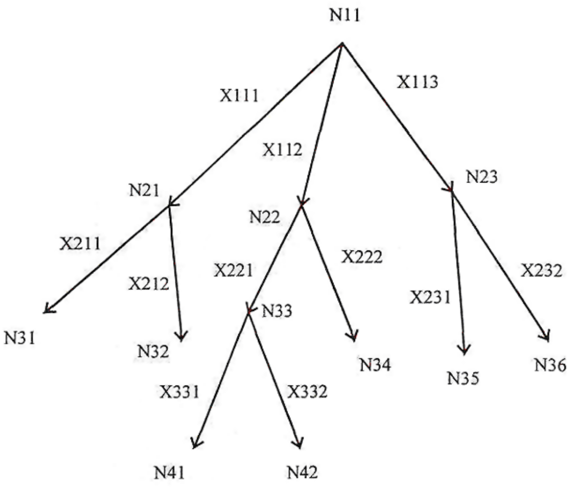

Figure 2. A graph representing decision tree

Let us denote nodes by Nij and branches by Xijk. Particular indices mean: i -level in the tree, j - number of node, k - number of branch. It will be noticed that we can convert the decision tree into a set of production rules. These rules have extended premises. They derive final conclusions without any intermediate steps. As an example the decision tree provided in Fig. 2 represents a set of 7 rules:

(N11, X111) AND (N21, X211) THEN N31
(N11, X111) AND (N21, X212) THEN N32
(N11, X112) AND (N22, X221) AND (N33, X331) THEN N41
(N11, X112) AND (N22, X221) AND (N33, X332) THEN N42
(N11, X112) AND (N22, X222) THEN N34
(N11, X113) AND (N23, X231) THEN N35
(N11, X113) AND (N23, X232) THEN N36

For the knowledge representation by the decision tree the appropriate rule may be selected by searching the tree. This process known as consultation with the user is some kind of a dialog. It should be noted that at each node of the decision tree except leafs value of attribute is established. Within the framework of this dialog the system asks the user for the facts needed to find values of attributes. All possible nodes are evaluated before proceeding to the next level of the decision tree. It causes a path to be formed among the nodes when we transverse the branches in only one direction. Each path emanating from the root and terminating at the leaf constitutes an inference path. During this process data input and output may vary a great deal from case to case. Note that our search is data driven, that is, we use the initial set of data to conduct the search.

Since the user answers synonymously, i.e. he chooses only one answer, at each node except leafs only one emanating branch may be selected. These branches create a set of input facts.

When consultation is performed in correct way there may be created only one inference path in given decision tree. This path enables identification of a leaf which represents an unknown conclusion. In conventional systems the user can input the correct answer on the screen using for example menu technology. Data can be also entered directly from the keyboard.

4 Encoding the decision tree by DNA molecules

In the molecular method described here the decision tree is encoded by DNA strands. Similarly as in [7] each node is encoded by the oligonucleotide of some length and sequence of nucleotides. The sequences are designed in such manner that strands representing nodes can **not** hybridize with each other. The strands representing nodes hybridize in unique way with strands representing adequate branches. Additionally, the oligonucleotide representing the root is labeled magnetically. Also the oligonucleotides representing leafs are labeled by quantum dots. Orientations of these strands are as depicted in Fig. 3.

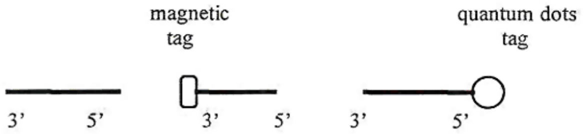

Figure 3. DNA strands for encoding the decision tree

Particular nodes as well as branches of the decision tree are encoded by single-strand oligonucleotides of specific sequence. The branch linking two nodes may be encoded in different way. For example length of the branch may be chosen to equal half of the sum of DNA lengths representing particular nodes. However other selections are also possible. Sequences of branch nucleotides should be complementary to sequences of respective oligonucleotides representing nodes.

Example

Suppose the DNA oligonucleotides representing nodes of the decision tree of Fig. 2 have been designed as follows:

N11 - root	5' CTAAATCCACTGTGATATC 3'
N21	5' TTGCTAAGCCAGCTGCTC 3'
N22	5' GTGCCTACCGAATCGCG 3'
N23	5' GGCTCAGCAATAGGCTCC 3'
N31 - leaf	5'CAACTGTTGCTGTTGAGG 3'
N32 - leaf	5'TCCGAATACGCCTA 3'
N33	5' ACAACGCCACTAGCGT 3'
N34 - leaf	5 'TTGCTCTGAGCATGCCGTGC 3'
N35 - leaf	5'CGAACGTTCCATAAGGGTCAC 3'
N36 - leaf	5 'TAGCGTAAGGTACGCAAACT 3'

N41 - leaf	5'TACCGAACGGTTGGCACGAT 3'
N42 - leaf	5' CTTCGAACGTTCCAG 3'

Let branches of the decision tree depicted in Fig. 2 be designed with following sequences:

X111	3' GATTTAGGTGACACTATAGAACGATTC 5'
X112	3' GATTTAGGTGAC ACT AT AGC ACGGATG 5'
X113	3' GATTTAGGTGACACTATAGCCGAGTCGTT 5'
X211	3' GGTCGACGAGGTTGACAACGACAACTCC 5'
X212	3' GGTCG ACGAGAGGCTTATGCGGAT 5'
X221	3' GCTTAGCGCTGTTGCGG 5'
X222	3' GCTTAGCGCAACGAGACTCGTACGGCACG 5'
X231	3' ATCCGAGGGCTTGCAAGGTATTCCCAGTG 5'
X232	3' ATCCGAGGATCGCATTCCATGCGTTTGA 5'
X331	3' TGATCGCAATGGCTTGCCAACCGTGCTA 5'
X332	3' TGATCGCAGAAGCTTGCAAGGTC 3'

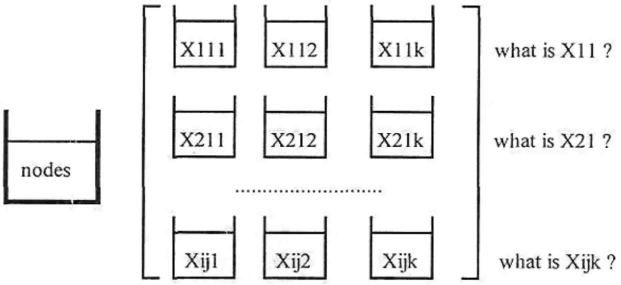

Figure 4. DNA water solutions encoding the decision tree

To encode the decision tree by molecules, we proceed in two steps. First, we prepare in a vessel the water solution of DNA which represents nodes of the decision tree. Next, we prepare a set of test tubes with water solutions of DNA which represent branches of the decision tree as provided in Fig. 4. During the consultation initial facts are established. The user selects branches emanating from nodes. At each node only one such branch may be chosen. This process is equivalent to selection of the proper tubes Xijk representing branches.

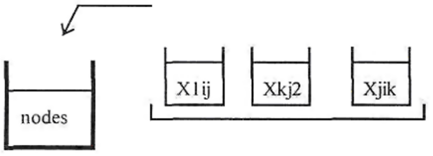

Figure 5. DNA water solutions after the consultation

When consultation is performed we get a set of tubes as depicted in Fig. 5. With respect to the structure of the decision tree and user answers offered during the consultation, there may be different sets of tubes Xijk.

5 Forming the inference path

To perform the inference on molecular level we should select proper DNA tubes representing branches of the graph as depicted in Fig. 5. Having mixed water solutions representing nodes and selected branches Xijk of the decision tree we examine results of biochemical reaction. Due to hybridization the oligonucleotides representing nodes and branches may connect with each other forming longer double-strand DNA chains. After ligation phospho-sugar backbones of particular strands are connected together forming permanent DNA molecules. Since proper inference paths begin with the root strand which is magnetically marked we select molecules with magnetic markers by purifying the solution. It can be done by creating a magnetic field affecting the vessel. Thus appropriate molecules will be retained during washing out the vessel. As a result only molecules with magnetic markers remain in the vessel while other molecules are removed.

Example

To explain the method suppose that after the consultation in graph of Fig. 2 there has been created the inference path as depicted by the continuous line in Fig. 6.

In this case during the consultation the following branches have been selected: X112, X221, X332. The inference path emanates from the root N11 and terminates at the leaf N42. It comprises two intermediate nodes N22 and N33. The leaf N42 represents final conclusion and it is objective of our inference.

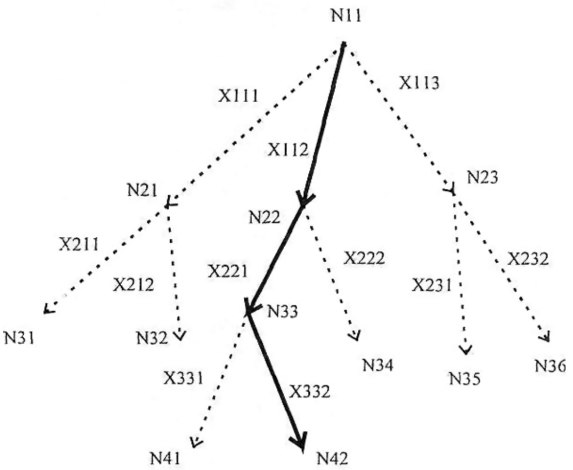

Figure 6. An example of the inference path in decision tree of Fig. 2

For the considered example of inference path (see Fig. 6) after hybridization, ligation and purification in final solution there may be 7 kinds of DNA molecules as presented in Fig. 7. These structures represent the possible chains created during hybridization of the seven DNA strands required for this inference path. Other connections of strands are not permitted because nucleotide sequences have been designed for unique hybridizations of strands representing nodes and branches. This is the key that locks adequacy between these two kinds of molecules. Therefore strands representing nodes can not hybridize with each other. This fact means that biochemical reaction is very selective comparing to ordinary chemical reactions.

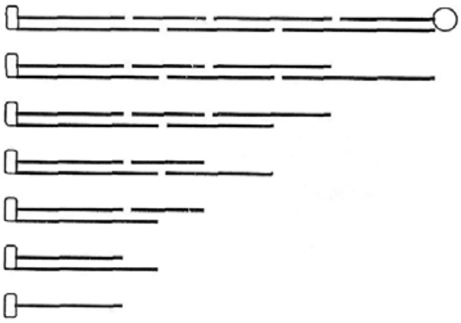

Figure 7. Possible chains of DNA after purification

As has been mentioned previously, lengths and nucleotide sequences of created
molecules depend on the structure of the decision tree as well as user answers during the consultation. If the DNA chain has the following properties: a) it starts with oligonucleotide representing the root, b) it passes through subsequent intermediate nodes, c) it terminates at the leaf, then such molecule constitutes an inference path.

As follows from Fig. 7 in our example six molecules have sticky ends. A single-strand molecule represents the root only. Five molecules are composed of strands representing root, branches, and intermediate nodes respectively. They do not terminate with a strand representing a leaf. Therefore these molecules do not represent inference path. However one of the molecules in Fig. 7 fulfills the requirements. The chain created by strands in this particular case is

{N11, X112, N22, X221, N33, X332, N42}

which provides the following sequence of nucleotides

5' CTAAATCCACTGTGATATCGTGCCTACCGAATCGCGACAACGCCACTAGCGT
CTTCGAACGTTCCAG 3'

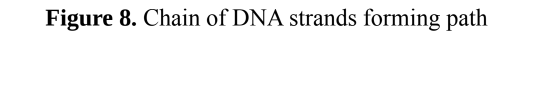

Figure 8. Chain of DNA strands forming path

This molecule is terminated with quantum dots tag as is shown in Fig. 8. In the last step of our procedure we detect the final result by using optical method as described in point 2. Since quantum dots allow the labeling a great number of molecules this approach seems to be very simple and fast.

6 Conclusion

We have presented a method of reasoning based on decision tree by manipulating DNA chains. Having prepared representation of nodes by adequate water solution of DNA, a set of initial facts is created by interacting with the user. During the consultation the user is asked a series of questions. Answers to these questions determine branches in the decision tree. In our approach these branches are represented by oligonucleotides in form of water solution to which are added the water solution of DNA representing the nodes. Hybridization, ligation and purification are performed by standard methods of genetic engineering. Finally, the inference path is detected by optical method using quantum dots. Result of reasoning depends on proper selection of branches during the consultation.

The fundamental difference between this method and that reported in [7] is the manner of identification of the leaves. In the previous method the length of any inference path is accounted by a number of nucleotide pairs in completely hybridized molecule. This length is determined by the oligonucleotide lengths representing particular nodes. Therefore, lengths of oligonucleotides representing nodes in the decision tree should be selected carefully so that each inference path would be of different length. There is no such restriction in the presented method.

In the previously reported approach, after hybridization and ligation, double-strand DNA chains were created as paths among the nodes of the decision tree. Then an autoradiography search for a DNA chain with an assigned radioactive oligonucleotide was performed. For radioactive molecules we determined their lengths by electrophorethical process. Detected DNA chain is considered as the inference path if its length belongs to a set of permitted values. Identified inference path indicates a conclusion encoded by the decision tree leaf. In the present approach the complicated and slow process of gel electrophoresis is eliminated.

Bibliography

[1] P. Węgleński, Molecular Genetics, (in Polish), PWN, Warsaw, 1995.

[2] J.J. Mulawka, Expert Systems, (in Polish), WNT, Warsaw, 1996.

[3] J.J.Mulawka, P. Borsuk, P. Węgleński, Implementation of the inference engine based on molecular computing technique, *Proc. of the IEEE International Conference on Evolutionary Computation*, Anchorage, 1998, 493 – 498.

[4] P. Wąsiewicz, T. Janczak, J.J. Mulawka, A. Płucienniczak, The inference based on molecular computing, Cybernetics and Systems Int. Journal, vol. 31, no. 3, 2000, 283-315.

[5] P. Wąsiewicz, T. Janczak, J.J. Mulawka, A. Płucienniczak, The inference via DNA computing, Proc. Congress on Evolutionary Computation, Washington, vol. 2, 1999, 988-993.

[6] J.J. Mulawka, M.J. Oćwieja, Molecular inference via unidirectional chemical reactions. Int. Conf. on Evolvable Systems: From Biology to Hardware, Lausanne, Lecture Notes in Computer Science, 1478, Springer, 1998, 372-379.

[7] J.J. Mulawka, T. Janczak, P. Borsuk, P. Węgleński, Reasoning via DNA based decision trees, Proc. Int. Conf. on Rough Sets and Current Trends in Computing, Warsaw 1998, 17-26.

[8] P.Wąsiewicz, J.J. Mulawka, Molecular genetic programming, Soft Computing, Springer, vol. 5, no. 2, 2001, 106-113.

[9] J.J. Mulawka, T. Janczak, A. Malinowski, R. Nowak, DNA computing - promise for information processing, Universitatis Jagellonicae Acta Informaticae, MMCCXLII, Z. 10, Krakow, 2000, 113-130.

[10] P. Wąsiewicz, J.J. Mulawka, Exceeding frontiers of microelectronics (in Polish), *Przegląd Telekomunikacyjny*, LXXIV, nr 1, 2001, 8-15.

[11] J.J. Mulawka, Molecular computing promise for new generation of computers, Proc. Polish-Czech-Hungarian Workshop on Circuit Theory, Signal Processing and Applications, Budapest, Wrzesień 1997, 94-99.

[12] L. Adleman, Molecular computation of solutions to combinatorial problems, Science, vol. 266, 1021-1024, 1994.

[13] R.J. Lipton, DNA solution of hard computational problems, Science, 268: April 28, 1995, 542-545.

[14] M. Ogihara, A. Ray, Simulating boolean circuits on a DNA computer, Technical Report TR 631, University of Rochester, Computer Science Department, August 1996.

[15] P. Wąsiewi.cz, A. Malinowski, R. Nowak, J.J. Mulawka, P. Borsuk, P. Węgleński, A. Płucienniczak, DNA computing implementation of data flow logical operations, Future Generation Computer Systems, Elsevier, 17, 2001, 361-378.

[16] J.J. Mulawka, P. Wąsiewicz, A. Płucienniczak, Logical operations with DNA strands, Proc. Int. Conf. on Rough Sets and Current Trends in Computing, Warsaw 1998, 27-36.

[17] P. ąsiewicz, P.Borsuk, J.J. Mulawka, P. Węgleński, Implementation of data flow logical operations via self-assembly of DNA, Lect. Notes in Computer Science, 1586, Springer, 1999, 174-182.

[18] E. B. Baum, Building an associative memory vastly larger than the brain. Science, 268, April 28, 1995, 583-585.

[19] E. Csuhaj-Varju, L. Kari, G. Paun, Test tube distributed systems based on splicing, Computers and AI, 15 (2-3), 1996, 211-232.

[20] A Bibliography of Molecular Computation and Splicing Systems, Ray Dassen http://liinwww.ira.uka.de/bibliography/Misc/dna.html.