Molecular binary trees

Grzegorz Tomczuk, Piotr Wąsiewicz
Institute of Electronic Systems
gtomczuk@elka.pw.edu.pl, pwasiewi@elka.pw.edu.pl

It seems that when in near future the potentiality of traditional semiconductor technology will probably be depleted, the nanotechnology and self-assembling feature of molecules will become the research main trend. The first step in this direction is so called molecular computing as the result of interference between computer science and genetic engineering. In this paper we propose new concepts of molecular binary trees.

1 Introduction

According to well-known Moore's law a number of transistors in the same area doubles every 18^{th} month. Further development work in electronic circuits miniaturization will be restricted by utilized in photolithography light wavelength and quantum effect problems. There is a need for some alternative technologies [1]. One possible solution is to carry out calculations on molecular level. Quantum computing and molecular computing are potential candidates and are under extensive development, recently. To perform computing different molecules in liquid and solid states may be used. Among them DNA acid has gained much attention.

In DNA computing [2] information is stored in molecules, which are linear polymers built from four building blocks - nucleotides denoted by symbols A, C, G, and T. Single or double DNA fragments are often called oligonucleotides or oligos, primers, strings and strands. In DNA computing a *string* is represented by a sequence of four basic nucleotides. DNA computing may be considered as a set of processing steps on DNA molecules for solving a specific problem according to a precisely defined procedure. A solution usually obtained by separating correct molecules from a set of all others representing the whole search solution space (placed e.g. in one vessel containing 10^{20} DNA strings) is reached by the exclusive use of genetic engineering operations on DNA such as hybridization, denaturation, ligation, PCR, etc. An executed operation supplies some DNA molecules as the result, which is identified usually by electrophoretical fluorescent or radioactive method. DNA computing methodology proposes models of massively parallel architectures and unique algorithms for them.

DNA computing is not only adequate for processing symbols and logical structures in general implementations of DNA computer [7], especially those solving NP-complete problems, but also for creating alternative computer architectures designed for molecular inference systems, molecular neural networks or evolutionary programming [3–6, 8]. Therefore molecular DNA computing is very close to computational intelligence.

2 Typical Operations on DNA Strings

According to the details of the biochemical structure and synthesis, DNA molecules are directional polymers. Their beginning is denoted as 5' and their end as 3'. Due to specific stereochemical interactions between A:T and C:G nucleotides DNA molecules can form antiparallel duplexes, provided that their sequence is complementary - allowing to form A:T and C:G pairs. Therefore in double stranded DNA the information is stored in both strands, in standard and complementary sequence. Transcription between standard

and complementary encoding is straightforward and is often used in the presented below algorithm.

It may exist as a separate DNA fragment or within a longer one e.g. a string a may be denoted by a sequence: $5'AGTC3'$ or may exist within a longer string $z = 5'AGA-AGTCCTA3'$. A formal language may be created based on DNA strings. The set of all single DNA strings over the alphabet $\Lambda = \{A, T, G, C\}$ is called the basic language of DNA computing and denoted by Λ^*.

A string *complementary* to, in this case, the standard string a is described by the same letter, but with an added symbol tilde (˜) or overline (-) this means $\tilde{a}$ or $\overline{a}$. Two complementary strings: a standard one a and a complementary one $\tilde{a}$ create after hybridization a double stranded string $\hat{a}$ made of complementary pairs: A with T and C with G.

Operations on DNA oligos [9] may be described in the following way:
1. *Hybridization or renaturation* means connecting or annealing of single complementary DNA strings to single standard DNA strings and forming double stranded molecules. This operation is caused by cooling down the test tube reaction solution and denoted by symbols heat $\downarrow$.
2. *Denaturation* means changing double stranded DNA molecules into single complementary and standard strings. Heating the test tube reaction solution causes it. Usually this operation is connected with the operation of *mixing* the solution. It is denoted by heat $\uparrow$.
3. *Cutting* of a double DNA string into two parts is performed in DNA computing with the help of enzymes. The resulting double strings can have sticky ends (single stranded DNA) or blunt ends.
4. *Concatenation* of two strings is a string formed by placing the second string after the first string without any gap. In DNA computing joining of two strings is done during hybridization and ligation. They form together a longer single string. In order to concatenate two oligos a and b the complementary to them in the place of joint, hybridized *third one* is needed.
5. *Amplification (PCR)*. In the cycle of amplification a single strings may be lengthened from its 3' end up to its complementary string 5' end e.g.:

$$\text{heat}\downarrow;\ \left\lceil \begin{matrix} [w &]^{\$} \\ \underline{-\quad v} \end{matrix} \right\rfloor \Rightarrow \left\lceil \begin{matrix} \tilde{v} \\ \underline{w} \end{matrix} \right\rfloor;\ \text{heat}\uparrow;$$

$$\text{heat}\downarrow;\ \left\lceil \begin{matrix} - & x \\ \underline{[\quad y]^{\$}} \end{matrix} \right\rfloor \Rightarrow \left\lceil \begin{matrix} \tilde{x} \\ \underline{x} \end{matrix} \right\rfloor;\ \text{heat}\uparrow;$$

$$\text{heat}\downarrow;\ \left\lceil \begin{matrix} \tilde{v} \\ \underline{w} \end{matrix} \right\rfloor \Leftarrow \left\lceil \left[\begin{matrix} + & w \\ \underline{\quad z \quad +} \end{matrix} \right]^{\$} \right\rfloor;\ \text{heat}\uparrow;$$

A sign + at the side of a DNA string describes a sticky end of it shorter than the nearest complementary oligo. A sign − at the right side of the DNA string describes a sticky end of it longer than the nearest complementary to it string. The same signs at both ends of complementary strings mean that these strings form a double stranded oligo with blunt ends. Note that the sign + may be additionally applied to mark a symbolic disjunction between two hybridized primers, and the sign $*$ to denote concatenation of strings (after hybridization and ligation), and the sign − to lengthen a string (of course, only in the equations). These rules are obligatory only within brackets $\lceil$ and $\rfloor$ or (and).
Every amplification described above is done in one cycle between cooling (heat $\downarrow$) and heating (heat $\uparrow$).
First of all, PCR encreases a number of double DNA strings chosen by specially designed primers two times in each cycle. The ends of these primers (square brackets) denote ends of amplified oligos. A number of PCR cycles is given in the upper, right corner of the right square bracket. If the number is unknown it is replaced by a sign \$. After tens of

amplification cycles in the test tube there are millions of chosen DNA fragments copies, which are in the majority. Amplification of double string can be described in another way as an algorithm:

$$\left\lceil \begin{array}{cccccccc} \tilde{a} & * & \tilde{e} & - & - & * & \tilde{b} \\ [& & & & p_2]^{\$} & & \\ & & [p_1 & & & &]^{\$} \\ \underline{a} & * & - & - & e & * & \underline{b} \end{array} \right\rceil \Rightarrow \left\lceil \begin{array}{cccccc} & & \tilde{e} & - & - & * & \tilde{b} \\ & & [& & p_2]^{\$} & & \\ & & [p_1 & &]^{\$} & & \\ \underline{a} & * & - & - & e & & \end{array} \right\rceil \Rightarrow \left\lceil \begin{array}{c} \tilde{e} \\ \underline{e} \end{array} \right\rceil;$$

where three amplification cycles are presented, and additional primers p_1, p_2 are short oligos complementary to small parts of the given double string. The special PCR enzyme builds from the hybridized primer a longer DNA string by adding to the primer 3' end nucleotides complementary to the string which this primer is attached to. The amplification result is in the double stranded form. The next step amplification is also performed on created in the previous steps assembled from nucleotides strings. Thus, primers are usually added in great quantities to the test tube before amplification, just in the order to be ready for hybridization to complementary parts and their successive lengthening during amplification.

6. *Mixing* of DNA fragments enables their uniform distribution. It improves search for good hybridizations in the space of all possible ones.

7. *Extracting* of DNA fragments with specific sequences among other DNA strings can be performed in several ways.

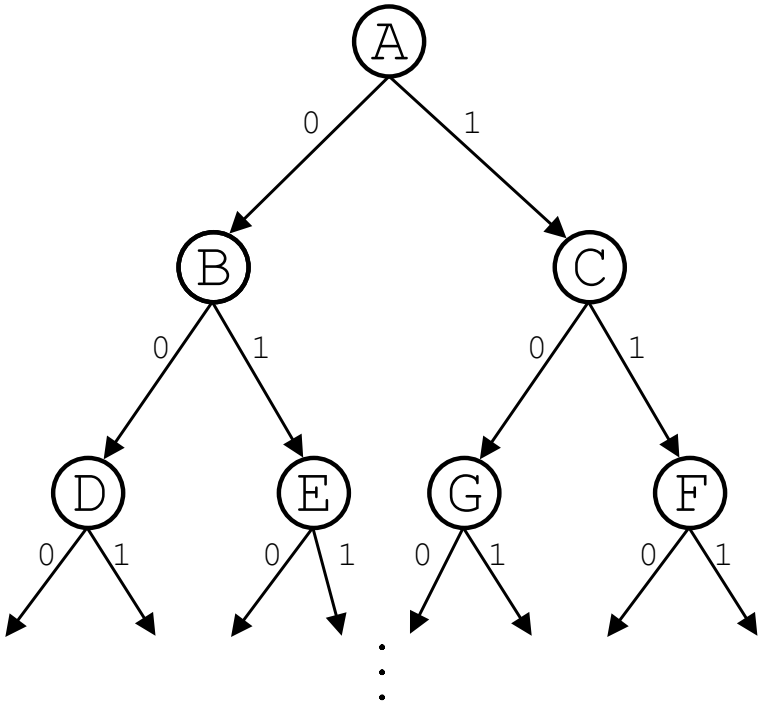

Fig. 1. Diagram of the binary tree modelled with DNA molecules.

3 Molecular tree model

Our algorithm of processing appropriate DNA molecules representing particular tree nodes allows identification of each tree node and additionally binary path for given answers equal to zeros or ones in particular node tests.

The basic element of the tree molecular model are specially designed DNA sequences of nodes. The first method depends on the whole tree from Fig. 1 encoding in one DNA string as is shown in Fig. 2.

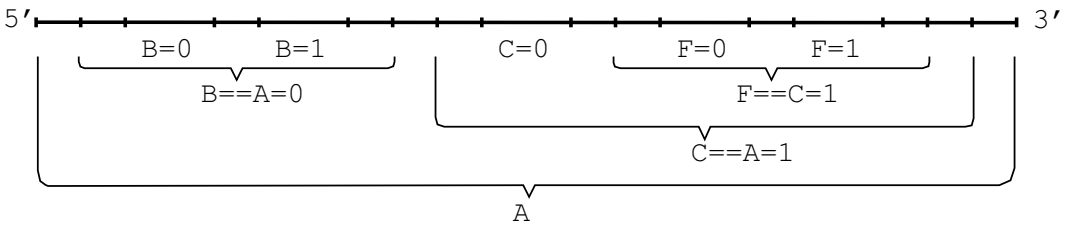

Fig. 2. Encoding of $A, B, C, .., F$ nodes in one DNA string.

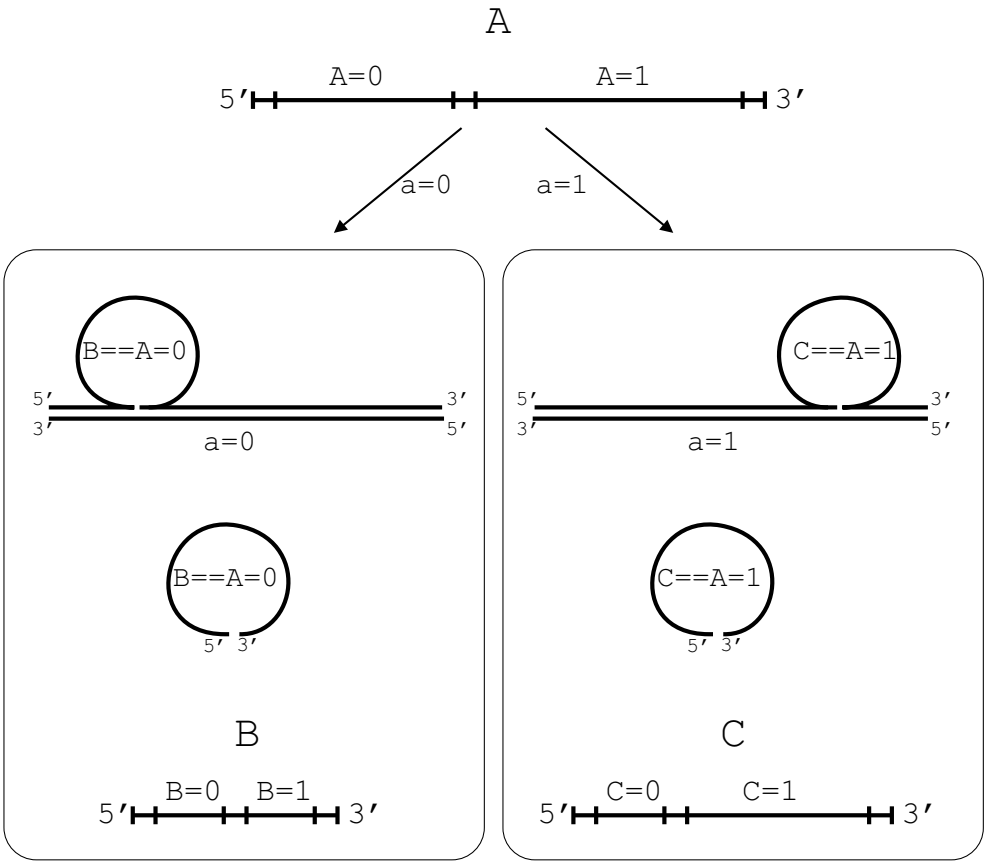

Fig. 3. Processing of node A depending on binary value connected to this nodeinput.

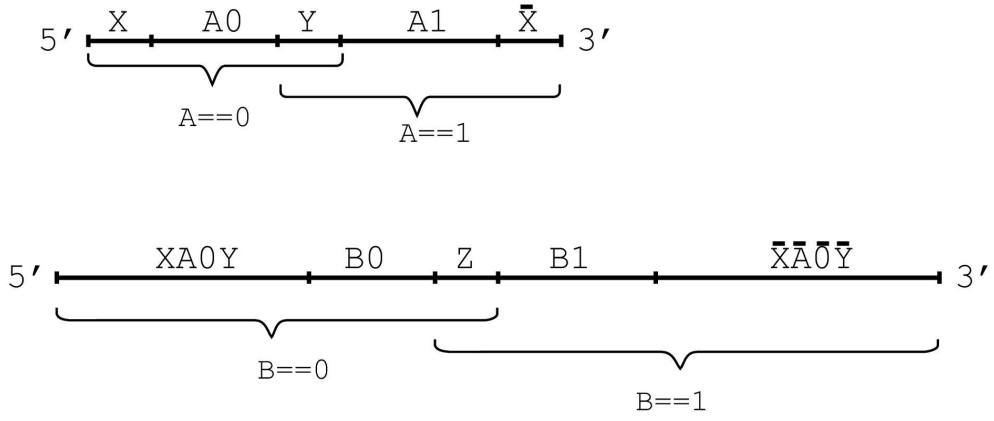

Fig. 4. Encoding of nodes A, B in two separate DNA strings

Processing of the node A starts after receiving input signal connected to this node. The corresponding complementary DNA fragments denoted by a representing the value

INPUT	INPUT MOLECULES	OUTPUT	OUTPUT MOLECULES
0	X, $\bar{Y}$	0	XA0Y
1	X, Y	1	YA1$\bar{X}$

TABLE I
The node A test signal encoding. The molecule X is needed in two cases.

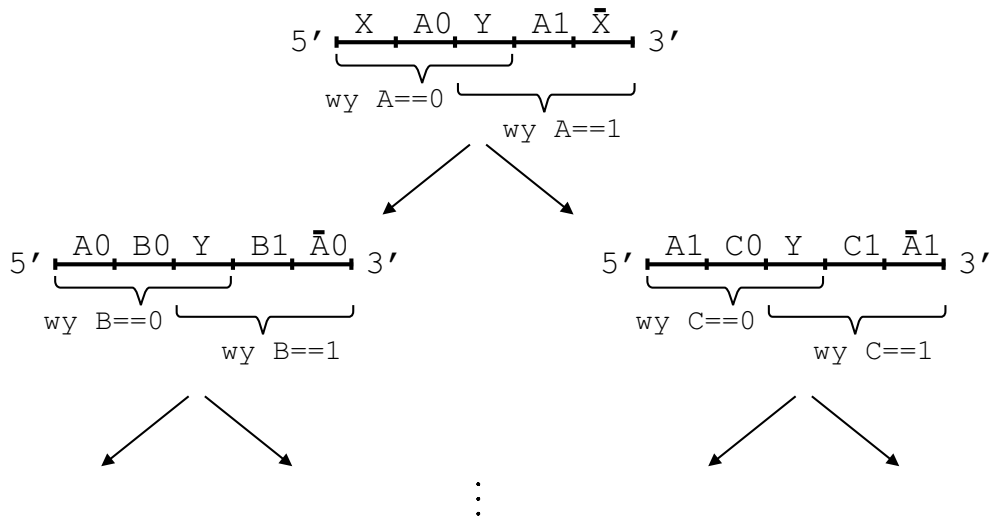

Fig. 5. Encoding of nodes A, B in two separate DNA strings without extending their ancestor lenghts

zero or one hybridize with one string tree and redundant branches are cut off by special enzymes destroying only one stranded (single) molecules. In Fig. 3 double molecules are the correct branches after processing value test in the node A. They are denatured and standard (not complementary) DNA strings are prepared for the next node value test and processing.

Another method of generating decision trees is depicted in Fig. 4. Every node is encoded in one unique molecule. In this string the left sector is connected with test value equal to zero and the right sector equal to one. The node A ends denoted by X and $\overline{X}$ are complementary to each other and prevent the amplification reaction of two sectors only with the help of X primers. An additional primer $\overline{Y}$ complementary to standard form of the node A two value molecule have to be included. It helps in the amplification process of the left sector connected with the test value equal to zero. Another additional primer Y complementary to complementary form of the node A two value molecule helps in the amplification process of the right sector connected with the test value equal to one. The input test answer data and evaluation results are written in Tab. I. The amplified sectors of molecule A will be ends of next node B and will prevent amplification of molecule B only with the help of primers complementary to these ends. The next primers Z and $\overline{(Z)}$ are needed for further evaluation of node B test.

With the present molecule $XA0Y$ it will be possible to generate answers during node B test. With the help of the rigth sector molecule $YA1X$ the node C test will be activated.

The method described in Fig. 5 is similar to the previous one where the succesive nodes needed longer and longer DNA strings, but here the ends complementary to primers e.g. X and Y are removed and the middle fragments of amplified solution representing directly each node are used as next test primers which after test evaluation are removed from the next test amplified solution. These middle fragments have the same lengths with unique sequences for each node. Thus, each node molecule has the same length and this method is more universal.

4 Summary

With proposed methods the whole decision trees can be generated on molecular scale. After adding appropriate answers for each node test the processing of decision path is activated and proceeds through successive node test value amplifications each after another.

Further research should extend ideas and give some approximation of graph self-assembled macromolecules.

Bibliography

[1] M.S. Malone, Beyond semiconductors, *The Microprocessor: A Biography*, Springer-Verlag (1995).

[2] *The Bibliography of Molecular Computation and Splicing Systems*, at http://-liinwww.ira.uka.de/bibliography/Misc/dna.html

[3] G. Tomczuk, P. Wąsiewicz, A. Pluciennicz ak. Molecular Neuron Network Experimental Approximation. Accepted for WSEAS Conference. Prague 2005.

[4] G. Tomczuk, P. Wąsiewicz, A. Dydynski, J.J. Mulawka, A. Pluciennicz ak. Molecular Neuron Realization. *WSEAS Transactions Journal on Biology and Biomedicine*, **1**(1):73-75.

[5] P. Wąsiewicz, T. Janczak, J.J. Mulawka, A. Płuciennicz ak, The Inference Based on Molecular Computing, *Cybernetics and Systems: An International Journal*, Taylor & Francis, vol. **31/3** (2000) 283-315.

[6] P. Wąsiewicz, A. Malinowski, R. Nowak, J.J Mulawka, P. Borsuk, P. Węgleński, A. Płuciennicz ak, DNA Computing: Implementation of Data Flow Logical Operations, *Future Generation Computer Systems Elsevier Journal*, **17/4** (2001) 361–378.

[7] P. Wasiewicz et al., Adding Numbers with DNA, *Proc. 2000 IEEE International Conference on Systems, Man & Cybernetics - SMC2000*, Nashville, USA (2000) 265-270.

[8] P. Wąsiewicz, J.J. Mulawka, Molecular Genetic Programming, *Journal of Soft Computing*, Springer, **5(2)** (2001).

[9] J. Sambrook, E.F. Fritsch, T. Maniatis, *Molecular Cloning. A Laboratory Manual.*, Second Edition, Cold Spring Harbor Laboratory Press (1989).