BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES
Vol. 52, No. 3, 2004

Resolving power of isothermic DNA sequencing chips

P. FORMANOWICZ[1,2]*

[1] Institute of Computing Science, Poznań University of Technology, 3A Piotrowo St., 60–965 Poznań, Poland
[2] Institute of Bioorganic Chemistry, Polish Academy of Sciences, 12/14 Noskowskiego St., 61–704 Poznań, Poland

Abstract. DNA sequencing remains one of the most important problems in molecular and computational biology. One of the methods used for this purpose is sequencing by hybridization. In this approach usually DNA chips composed of a full library of oligonucleotides of a given length are used, but in principle it is possible to use another types of chips. Isothermic DNA chips, being one of them, when used for sequencing may reduce hybridization error rate. However, it was not clear if a number of errors following from subsequence repetitions is also reduced in this case. In this paper a method for estimating resolving power of isothermic DNA chips is described which allows for a comparison of such chips and the classical ones. The analysis of the resolving power shows that the probability of sequencing errors caused by subsequence repetitions is greater in the case of isothermic chips in comparison to their classical counterparts of a similar cardinality. This result suggests that isothermic chips should be chosen carefully since in some cases they may not give better results than the classical ones.

Keywords: computational biology, DNA sequencing, DNA chips, sequencing errors, resolving power.

1. Introduction

One of the most important problems in molecular and computational biology is reading of DNA sequences. As it is known for over fifty years the genetic information determining structure and functionality of all living organisms is encoded in DNA molecules (some viruses are exceptions to this rule but it is a philosophical question whether are they living organisms).

DNA molecule is composed of four types of basic components called nucleotides, which are denoted by A, C, G, and T. A sequence of nucleotides composing DNA molecule determines the genetic information encoded in it. From this follows that in order to read the information it is necessary to determine the sequence of nucleotides. Unfortunately, it is impossible to read such a sequence by some direct method, e.g. using a microscope, because the structure of the molecule is too complex. Instead, some indirect methods are used. One of them is sequencing by hybridization (SBH) [1–3]. This method consists of two stages: the biochemical one and the computational one.

In the former, biochemical stage it is determined which sequences from some predefined set of short sequences the examined DNA sequence contains as a subsequence. In the classical variant of the method this set consists of all sequences of a given length l. To this set there corresponds a collection of short single stranded DNA molecules called *oligonucleotides*. Such a collection is named an *oligonucleotide library*. Each element of the set is represented by a number of oligonucleotides in the library. Nowadays, the library is made as a *DNA chip* [4–7]. Such a chip is a matrix divided into a number of cells each containing (in the classical version of the chip) all oligonucleotides of a given type being elements of the library. These oligonucleotides are attached to the surface of the chip.

The possibility of using DNA chips in the sequencing process follows from one of the fundamental properties of nucleic acids (DNA being one of them), i.e. their ability to create double stranded complexes. This means that two single stranded DNA molecules create a duplex if in these molecules there are pairs of complementary nucleotides. To be more precise, nucleotide A is complementary to nucleotide T, and nucleotide C is complementary to nucleotide G. If in two single stranded DNA molecules there are subsequences of complementary nucleotides, the sequences will join by hydrogen bonds, i.e. they will hybridize to each other. Moreover, it should be said that a pair of nucleotides A and T will be joined by two hydrogen bonds, while a pair C and G, by three such bonds. For example, if one of the single stranded DNAs contains a subsequence ATCAGT, and the other one contains TAGTCA, they can hybridize, since the subsequences are complementary. The rule is known as Watson-Crick complementarity rule [8].

The DNA chip when put into a solution of a number of copies of the examined single stranded DNA molecule may serve as a detector of l-tuples being subsequences of these molecules. If in the target sequence there is a subsequence complementary to oligonucleotides located in one of the cells of the chip such a molecule may hybridize to the oligonucleotide. Since in the solution there may be hundreds of thousands or millions of target DNA strands a number of them will hybridize to those oligonucleotides attached to the chip which are complementary to some subsequence of the strand. The examined molecules may be fluorescently or radioactively labelled, hence by analysing the image of the chip it is possible to get information where the molecules have hybridized. This information is equivalent to the one concerning the l-tuples composition of the target DNA. However, the information is not sufficient for the discovery of the nucleotide sequence of the target DNA since the order of the l-tuples remains unknown.

* e-mail: piotr@cs.put.poznan.pl

That is the end of the first stage of the SBH method. In the latter, computational one, on the base of the set of l-tuples obtained in the former stage (the set is called *spectrum*) the target sequence is recovered. (Note, that the cardinality of the spectrum is equal to $n-l+1$, if n is the length of the target sequence.)

The biochemical phase of the SBH method described above is in fact its ideal variant where no experimental errors occur. In the real world experiments two main types of errors are involved in the resulting data [9]. One of them is called *negative errors* or *false negatives*. Such errors follow from situations where the target DNA should hybridize to oligonucleotides on the chip (because it is 100% complementary to them) but it has not. In this case the spectrum is incomplete — information on some l-tuples composing the target molecule has been missed. It should be stated that false negatives will also occur when the hybridization reaction is perfect, but the target sequence contains repetitions of some l-tuples. In such a case spectrum will contain information about the presence of such l-tuples but it will not contain any information concerning multiplicity of the l-tuples (which follows from the current technology used in the biochemical phase). So, in this case spectrum contains information about all types of l-tuples composing the target molecule, but not about all l-tuples (in this case its cardinality is less than $n-l+1$).

On the other hand, it may happen that the target DNA molecule will hybridize to some oligonucleotides which are not 100% complementary to it. In such a case spectrum will contain false information about some l-tuples which are not part of the examined sequence (its cardinality is greater than $n-l+1$). That is a source of *positive errors* or *false positives*. In practice both types of errors occur in most of the experiments.

In general, it is possible to reduce the error rate by using different types of oligonucleotide libraries. One of the possible variants of the library is an *isothermic* one [10].

In the paper in Section 2. the isothermic oligonucleotide libraries will be described. In Section 3. resolving power of the isothermic chips, i.e. chips composed of isothermic oligonucleotide libraries will be derived while in Section 4. the results of a computational analysis will be shown. The paper ends with conclusions in Section 5.

2. Isothermic DNA chips

The method described in the previous section is a classical variant of SBH, where DNA chips used are composed of a full library of all l-long oligonucleotides for a given l. As has been mentioned in the previous section chips of other types can be used which may lead to a reduction of an error rate and one type of such chips are the isothermic ones.

Roughly speaking an isothermic chip is composed of all oligonucleotides which creates duplexes of a given melting temperature. This temperature is equivalent to energy of the duplex bonds. The calculation of an exact value of the temperature is a complicated thermodynamic problem but some simplified models also exist. In one of them each pair A–T adds 2 deg to the melting temperature of the DNA complex while each C–G pair adds 4 deg [11]. So, in this model a nucleotide composition (but not the sequence of the nucleotides) of a given oligonucleotide determines its melting temperature.

More formally, the isothermic oligonucleotide library is a special case of *isorelational oligonucleotide library* defined as follows [10]:

Definition 1. An isorelational oligonucleotide library is a library L consisting of all oligonucleotides satisfying relation $w_A x_A + w_C x_C + w_G x_G + w_T x_T = C_L$, where w_A, w_C, w_G, w_T are increments of nucleotides A, C, G and T, respectively, x_A, x_C, x_G, x_T denote numbers of these nucleotides in the oligonucleotide, and C_L is a constant parameter for the library.

Based on the above definition an isothermic oligonucleotide library can be defined as follows [10]:

Definition 2. An isothermic oligonucleotide library L of temperature τ_L is a library of all oligonucleotides satisfying relations $w_A x_A + w_C x_C + w_G x_G + w_T x_T = \tau_L$, $w_A = w_T$, $w_C = w_G$ and $2w_A = w_C$.

Without loss of generality it may be assumed that $w_A = w_T = 2$ and $w_C = w_G = 4$. This corresponds to the amount of energy or equivalently temperature which nucleotides bring into the stability of oligonucleotide duplexes. The most important property of isothermic library is its ability to form duplexes in a more narrow range of experimental conditions (i.e. temperature, salt concentration) than in the case of classical libraries. It means that in a given conditions all duplexes created by elements of isothermic library have approximately the same stability (i.e. the energy of their hydrogen bonds has approximately the same value), which leads to a reduction of hybridization error rate.

Another important feature of the isothermic libraries is the fact that one such a library is not sufficient for DNA sequencing but two such libraries with melting temperatures differing by 2 deg can be used for sequencing any DNA molecule [10].

Other important properties of the libraries are their cardinalities which may be calculated according to the following formulae, where (1) concerns libraries with temperature τ divisible by 4, while (2) concerns libraries with τ non-divisible by 4.

$$\mathrm{card}(\tau) = \sum_{i=0}^{\frac{\tau}{4}} \left[\binom{\frac{\tau}{4}+i}{2i} 2^{\frac{\tau}{4}+i} \right] \tag{1}$$

$$\mathrm{card}(\tau) = \sum_{i=0}^{\lfloor \frac{\tau}{4} \rfloor} \left[\binom{\lfloor \frac{\tau}{4} \rfloor + i + 1}{2i+1} 2^{\lfloor \frac{\tau}{4} \rfloor + i + 1} \right] \tag{2}$$

It is easy to notice that these cardinalities are between the cardinalities of the standard libraries involving, respectively, only the shortest ($l = \frac{\tau}{4}$ or $l = \lceil \frac{\tau}{4} \rceil$) or only the longest ($l = \frac{\tau}{2}$) oligonucleotides of equal length.

3. Branching probability

As has been said in the previous Sections in most of the approaches to SBH classical DNA chips are used, i.e. chips containing a full library of all 4^l oligonucleotides of length l. On the other hand, it is known for some time, that in principle it is possible to design many other DNA chips which may be used in DNA sequencing [7,12,13]. From this fact there emerges a need to have some measure which might be useful in comparing different chip designs and answer the question which one is the best one for DNA sequencing. Such a measure has been proposed and an analysis of some non-classical chips have been shown in [7].

Below we show such an analysis for classical chips (it is the analysis made in [7] with some details additionally explained) and we analyze in a similar way isothermic DNA chips.

Let us consider DNA sequence $Q = X_1X_2 \ldots X_{m-1} X_mX_{m+1} \ldots X_n$, i.e. a sequence over alphabet $\Sigma_{DNA} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ and assume that its prefix $Q_m = X_1X_2 \ldots \ldots X_m$ has already been determined.

The goal is to estimate the probability of ambiguous extending sequence Q_m to the right by one nucleotide in the case of using isothermic DNA chips.

Let us denote by $C_{it}(\tau, \tau+2)$ an isothermic DNA chip composed of oligonucleotide libraries of melting temperatures equal to τ and $\tau + 2$. In what follows a cell on the chip will be called a *probe*.
Let

$$S(C,Q) = \left\{ \begin{array}{l} p \in C : \text{probe } p \text{ contains oligonu-} \\ \text{cleotide complementary to some} \\ \text{subsequence of sequence } Q. \end{array} \right\} \quad (3)$$

Let us observe that

$$S(C, Q_m) \subseteq S(C, Q). \quad (4)$$

Sequence Q_m may be extended by one nucleotide in four ways, i.e. $Q_m\,\mathrm{A}$, $Q_m\,\mathrm{C}$, $Q_m\,\mathrm{G}$, and $Q_m\,\mathrm{T}$.

The extension of sequence Q_m by nucleotide $\eta \in \Sigma_{DNA}$ is a *feasible extension*, if

$$S(C, Q_m\eta) \subseteq S(C, Q). \quad (5)$$

Let

$$\omega(C,Q,m) = \left\{ \begin{array}{ll} 0 & \text{if } S(C, Q_m\eta) \subseteq S(C,Q) \text{ holds} \\ & \text{for exactly one nucleotide} \\ 1 & \text{otherwise} \end{array} \right. . \quad (6)$$

We say that Q is *uniquely extendable* after m with respect to chip C if $\omega(C, Q, m) = 0$; otherwise we say that Q is *non-uniquely extendable*.

The branching probability $p(C, n, m)$ is the probability that a random sequence of length n is non-uniquely extendable after m-th nucleotide using chip C. This probability is equal to [7]

$$p(C,n,m) = \frac{1}{4^n} \sum_Q \omega(C,Q,m) \quad (7)$$

where the sum is taken over all sequences over alphabet Σ_{DNA} of length n.
We can fix m and denote $p(C, n) = p(C, n, m)$.

In the case of classical chip C_l (i.e. the one composed of all oligonucleotides of length l) it suffices to check if any of sequences $V\gamma_1, V\gamma_2, V\gamma_3$ belongs to $S(C_l, Q)$, where V is a suffix of length $l-1$ of sequence Q_m, and $\gamma_1, \gamma_2, \gamma_3$ are nucleotides different from X_{m+1}, i.e. in that case we say that sequence Q is non-uniquely extendable if $S(C_l, Q)$ contains $V\gamma_1$ or $V\gamma_2$ or $V\gamma_3$ [7].

In the case of isothermic chip the situation is a bit more complicated, because the extension of Q_m by nucleotide N may cause that $S(C_{it}(\tau, \tau+2), Q_m\,\mathrm{N})$ will include oligonucleotides of various lengths $l \in [\lceil \frac{\tau}{4} \rceil, \frac{\tau}{2}]$, which belong to $S(C_{it}(\tau, \tau+2), Q)$.

For a given l the probability of finding a given subsequence of length l at a given position in sequence Q is equal to

$$\frac{1}{4^l}. \quad (8)$$

The probability that this subsequence is not present at that position equals to

$$1 - \frac{1}{4^l}. \quad (9)$$

The probability that this subsequence is not present at any position of Q equals

$$\left(1 - \frac{1}{4^l}\right)^{n-l+1}. \quad (10)$$

Since our goal is to determine the probability of not finding any of the three subsequences $V\gamma_1$, $V\gamma_2$ and $V\gamma_3$ we obtain

$$\left(\left(1 - \frac{1}{4^l}\right)^{n-l+1}\right)^3. \quad (11)$$

The branching probability, i.e. the probability of finding at least one of those subsequences in Q is equal to

$$p_{br}(C_l, n) = 1 - \left(\left(1 - \frac{1}{4^l}\right)^{n-l+1}\right)^3. \quad (12)$$

The above formula describes the branching probability of a classical DNA chip containing oligonucleotides of length l [7].

In the case of isothermic DNA chips we have to take into account two additional facts:

1) on the chip there are oligonucleotides of various lengths,
2) for given l there is a proper subset of all 4^l oligonucleotides of length l on the chip.

Hence, considering the possibility of ambiguous extending sequence Q at a given position m we should consider the probability of two events:

B — at position m there ends subsequence $R\eta_1$, where $R \in \Sigma_{DNA}^{l-1}$ and $\eta_1 \in \Sigma_{DNA}$ of length l and temperature τ or $\tau+2$,

A — in sequence Q at position different form m there ends subsequence $R\eta_2$ of temperature τ or $\tau+2$, where $\eta_2 \in \Sigma_{DNA}$ and $\eta_2 \neq \eta_1$.

So, the branching probability can be expressed as

$$P_{br} = P(A)P(B) = (1 - P(\overline{A}))P(B). \tag{13}$$

For given l and τ we have:

$$P_{br}(l,\tau) = \left(1 - \left(1 - \frac{1}{4^l}\right)^{n-l+1}\right)\frac{N_\tau(l)}{4^l} \tag{14}$$

where $N_\tau(l)$ is a number of oligonucleotides of length l and temperature τ.

But, it is necessary to take into account all oligonucleotides' lengths for a given τ hence, we have the following:

$$P_{br}(\tau) = \sum_{l=\lceil \frac{\tau}{4} \rceil}^{\frac{\tau}{2}} \left[\left(1 - \left(1 - \frac{1}{4^l}\right)^{n-l+1}\right)\frac{N_\tau(l)}{4^l}\right]. \tag{15}$$

Moreover, we should take into account, that the chip is composed of two libraries of temperatures τ and $\tau+2$. Hence, we should add additional component to the previous formula:

$$\begin{aligned} P_{br}(\tau,\tau+2) &= \sum_{l=\lceil \frac{\tau}{4} \rceil}^{\frac{\tau}{2}} \left[\left(1 - \left(1 - \frac{1}{4^l}\right)^{n-l+1}\right)\frac{N_\tau(l)}{4^l}\right] \\ &+ \sum_{l=\lceil \frac{\tau+2}{4} \rceil}^{\frac{\tau+2}{2}} \left[\left(1 - \left(1 - \frac{1}{4^l}\right)^{n-l+1}\right)\frac{N_{\tau+2}(l)}{4^l}\right]. \end{aligned} \tag{16}$$

Formula (16) only approximates the real value of the branching probability. The first part of it describes the probability that at a given position in a random sequence there ends a subsequence of temperature τ and somewhere else in the sequence there is a similar subsequence that differs from the previous one only by the last nucleotide, but both of them have melting temperature equal to τ. The second part of the formula describes an analogous probability for subsequences of temperature $\tau+2$.

However, it may happen that in a random sequence Q at position m there ends subsequence $R\alpha$ of temperature τ, where $\alpha \in \{\mathrm{A},\mathrm{T}\}$ and at some other position in Q there ends subsequence $R\beta$, where $\beta \in \{\mathrm{C},\mathrm{G}\}$. Obviously, $R\beta$ has temperature $\tau+2$ and will appear in spectrum for sequence Q, but this situation is not taken into account by the first part of (16), neither by the second part. So, we should divide the two components in formula (16) into parts describing branching probability caused by subsequence of the same temperature as of this one found at position m, and by subsequence complementary to some oligonucleotide from the second library on the chip:

$$\begin{aligned} &P_{br}(\tau,\tau+2) \\ &= \sum_{l=\lceil \frac{\tau}{4} \rceil}^{\frac{\tau}{2}} \left[\left(1 - \left(1 - \frac{1}{4^l}\right)^{n-l+1}\right)\frac{N_{\tau(C/G)}(l)}{4^l}\right] \\ &+ \sum_{l=\lceil \frac{\tau}{4} \rceil}^{\frac{\tau}{2}} \left[\left(1 - \left(\left(1 - \frac{1}{4^l}\right)^{n-l+1}\right)^3\right)\frac{N_{\tau(A/T)}(l)}{4^l}\right] \\ &+ \sum_{l=\lceil \frac{\tau+2}{4} \rceil}^{\frac{\tau+2}{2}} \left[\left(1 - \left(1 - \frac{1}{4^l}\right)^{n-l+1}\right)\frac{N_{\tau+2(A/T)}(l)}{4^l}\right] \\ &+ \sum_{l=\lceil \frac{\tau+2}{4} \rceil}^{\frac{\tau+2}{2}} \left[\left(1 - \left(\left(1 - \frac{1}{4^l}\right)^{n-l+1}\right)^3\right)\frac{N_{\tau+2(C/G)}(l)}{4^l}\right] \end{aligned} \tag{17}$$

where: $N_{\tau(C/G)}(l)$ is a number of oligonucleotides of temperature τ and length l which have C or G at the last position, $N_{\tau(A/T)}(l)$ is a number of oligonucleotides of temperature τ and length l which have A or T at the last position.

Let us observe that the number of oligonucleotides of temperature τ divisible by 4 which have C or G at the last position is equal to

$$\sum_{i=0}^{\frac{\tau}{4}-1} \binom{\frac{\tau}{4}-1+i}{2i} 2^{\frac{\tau}{4}+i}. \tag{18}$$

The number of such oligonucleotides which end with A or T equals

$$\sum_{i=1}^{\frac{\tau}{4}} \binom{\frac{\tau}{4}-1+i}{2i-1} 2^{\frac{\tau}{4}+i}. \tag{19}$$

Moreover, for τ non-divisible by 4 the number of oligonucleotides which end with C or G is

$$\sum_{i=0}^{\lfloor \frac{\tau}{4}-1 \rfloor} \binom{\lfloor \frac{\tau}{4} \rfloor + i}{2i+1} 2^{\lfloor \frac{\tau}{4} \rfloor + i + 1}. \tag{20}$$

The number of such oligonucleotides which end with A or T is

$$\sum_{i=0}^{\lfloor \frac{\tau}{4} \rfloor} \binom{\lfloor \frac{\tau}{4} \rfloor + i}{2i} 2^{\lfloor \frac{\tau}{4} \rfloor + i + 1}. \tag{21}$$

From (18)–(21) it follows that for τ divisible by 4 we have:

$$N_{\tau(C/G)}^{even}(l) = \binom{l-1}{2(l-\frac{\tau}{4})} 2^l \tag{22}$$

and

$$N_{\tau(A/T)}^{even}(l) = \binom{l-1}{2(l-\frac{\tau}{4})-1} 2^l. \tag{23}$$

For τ non-divisible by 4 we have:

$$N_{\tau(C/G)}^{odd}(l) = \binom{l-1}{2(l-\lfloor \frac{\tau}{4} \rfloor)-1} 2^l \tag{24}$$

and

$$N^{odd}_{\tau(A/T)}(l) = \binom{l-1}{2(l-\lfloor\frac{\tau}{4}\rfloor)-2} 2^l. \tag{25}$$

Finally, from (17) and (22)–(25) for τ divisible by 4 we have (26), and for τ non-divisible by 4 we have (27). Equations (26) and (27) are shown at the bottom of the page.

Obviously, the lower the value of the branching probability the better. In the next Section formulae (26) and (27) are used to calculate the values of branching probability of isothermic chips having various melting temperatures.

4. Computational analysis

In this Section the values of branching probability for classical and isothermic DNA chips calculated according to the formulae derived in the previous Section are presented.

In Tables 1 and 2 branching probabilities for isothermic chips of temperatures τ and $\tau+2$, where τ is in the range from 26 to 40 deg are shown. These probabilities

Table 1

Branching probability for DNA chips composed of libraries of temperatures τ and $\tau+2$, where τ is in the range from 26 to 32 deg and target sequences of lengths in the range from 100 to 1000

sequence length	temperature τ			
n	26	28	30	32
100	0.002753	0.000808	0.000486	0.000141
200	0.005680	0.001680	0.001013	0.000295
300	0.008582	0.002549	0.001539	0.000449
400	0.011460	0.003416	0.002064	0.000603
500	0.014314	0.004281	0.002588	0.000757
600	0.017144	0.005143	0.003112	0.000910
700	0.019951	0.006004	0.003634	0.001064
800	0.022735	0.006862	0.004155	0.001218
900	0.025496	0.007717	0.004675	0.001371
1000	0.028234	0.008571	0.005195	0.001525

$$\begin{aligned}
p_{br}(C_{it}(\tau,\tau+2),n) = {} & \sum_{l=\lceil\frac{\tau}{4}\rceil}^{\frac{\tau}{2}} \left[\left(1-\left(1-\frac{1}{4^l}\right)^{n-l+1}\right)\frac{\binom{l-1}{2(l-\frac{\tau}{4})}2^l}{4^l}\right] \\
& + \sum_{l=\lceil\frac{\tau}{4}\rceil}^{\frac{\tau}{2}} \left[\left(1-\left(\left(1-\frac{1}{4^l}\right)^{n-l+1}\right)^3\right)\frac{\binom{l-1}{2(l-\frac{\tau}{4})-1}2^l}{4^l}\right] \\
& + \sum_{l=\lceil\frac{\tau+2}{4}\rceil}^{\frac{\tau+2}{2}} \left[\left(1-\left(1-\frac{1}{4^l}\right)^{n-l+1}\right)\frac{\binom{l-1}{2(l-\lfloor\frac{\tau+2}{4}\rfloor)-2}2^l}{4^l}\right] \\
& + \sum_{l=\lceil\frac{\tau+2}{4}\rceil}^{\frac{\tau+2}{2}} \left[\left(1-\left(\left(1-\frac{1}{4^l}\right)^{n-l+1}\right)^3\right)\frac{\binom{l-1}{2(l-\lfloor\frac{\tau+2}{4}\rfloor)-1}2^l}{4^l}\right]
\end{aligned} \tag{26}$$

$$\begin{aligned}
p_{br}(C_{it}(\tau,\tau+2),n) = {} & \sum_{l=\lceil\frac{\tau}{4}\rceil}^{\frac{\tau}{2}} \left[\left(1-\left(1-\frac{1}{4^l}\right)^{n-l+1}\right)\frac{\binom{l-1}{2(l-\lfloor\frac{\tau}{4}\rfloor)-1}2^l}{4^l}\right] \\
& + \sum_{l=\lceil\frac{\tau}{4}\rceil}^{\frac{\tau}{2}} \left[\left(1-\left(\left(1-\frac{1}{4^l}\right)^{n-l+1}\right)^3\right)\frac{\binom{l-1}{2(l-\lfloor\frac{\tau}{4}\rfloor)-2}2^l}{4^l}\right] \\
& + \sum_{l=\lceil\frac{\tau+2}{4}\rceil}^{\frac{\tau+2}{2}} \left[\left(1-\left(1-\frac{1}{4^l}\right)^{n-l+1}\right)\frac{\binom{l-1}{2(l-\frac{\tau+2}{4}-1)}2^l}{4^l}\right] \\
& + \sum_{l=\lceil\frac{\tau+2}{4}\rceil}^{\frac{\tau+2}{2}} \left[\left(1-\left(\left(1-\frac{1}{4^l}\right)^{n-l+1}\right)^3\right)\frac{\binom{l-1}{2(l-\frac{\tau}{4})}2^l}{4^l}\right].
\end{aligned} \tag{27}$$

Table 2

Branching probability for DNA chips composed of libraries of temperatures τ and $\tau + 2$, where τ is in the range from 34 to 40 deg and target sequences of lengths in the range from 100 to 1000

sequence length	temperature τ			
n	34	36	38	40
100	0.000086	0.000025	0.000015	0.000004
200	0.000180	0.000052	0.000032	0.000009
300	0.000273	0.000079	0.000048	0.000014
400	0.000367	0.000106	0.000065	0.000019
500	0.000461	0.000134	0.000082	0.000024
600	0.000555	0.000161	0.000099	0.000028
700	0.000649	0.000188	0.000115	0.000033
800	0.000743	0.000215	0.000132	0.000038
900	0.000836	0.000243	0.000149	0.000043
1000	0.000930	0.000270	0.000165	0.000048

are calculated for the target DNA sequences of lengths from 100 to 1,000 nucleotides. Obviously, when the length of the target sequence increases the value of the branching probability also increases. On the other hand, increasing the temperature of the isothermic library decreases the value of the probability. This observation is also not surprising since the probability of finding repetitions of a given subsequence in a random sequence decreases when the length of the subsequence increases.

In Tables 3 and 4 the values of branching probability for classical chips composed of oligonucleotides of lengths in the range from 8 to 22 are shown.

Obviously, in order to compare the values of branching probability of different types of chips it is necessary to take into account chips composed of similar numbers of oligonucleotides. The cardinalities of isothermic and classical chips are shown in Tables 5 and 6.

Table 3

Branching probability for DNA chips composed of full libraries of oligonucleotides of length l in the range from 8 to 14 and target sequences of lengths in the range from 100 to 1000

sequence length	oligonucleotide length l			
n	8	10	12	14
100	0.004248	0.000260	0.000016	0.000001
200	0.008796	0.000546	0.000034	0.000002
300	0.013323	0.000832	0.000052	0.000003
400	0.017829	0.001118	0.000070	0.000004
500	0.022315	0.001404	0.000087	0.000005
600	0.026780	0.001689	0.000105	0.000007
700	0.031225	0.001975	0.000123	0.000008
800	0.035650	0.002261	0.000141	0.000009
900	0.040054	0.002546	0.000159	0.000010
1000	0.044439	0.002831	0.000177	0.000011

Table 4

Branching probability for DNA chips composed of full libraries of oligonucleotides of length l in the range from 16 to 22 and target sequences of lengths in the range from 100 to 1000

sequence length	oligonucleotide length l			
n	16	18	20	22
100	$5.937 \cdot 10^{-8}$	$3.623 \cdot 10^{-9}$	$2.210 \cdot 10^{-10}$	$1.347 \cdot 10^{-11}$
200	$1.292 \cdot 10^{-7}$	$7.989 \cdot 10^{-9}$	$4.939 \cdot 10^{-10}$	$3.052 \cdot 10^{-11}$
300	$1.991 \cdot 10^{-7}$	$1.235 \cdot 10^{-8}$	$7.667 \cdot 10^{-10}$	$4.758 \cdot 10^{-11}$
400	$2.689 \cdot 10^{-7}$	$1.672 \cdot 10^{-8}$	$1.040 \cdot 10^{-9}$	$6.463 \cdot 10^{-11}$
500	$3.388 \cdot 10^{-7}$	$2.109 \cdot 10^{-8}$	$1.312 \cdot 10^{-9}$	$8.168 \cdot 10^{-11}$
600	$4.086 \cdot 10^{-7}$	$2.545 \cdot 10^{-8}$	$1.585 \cdot 10^{-9}$	$9.874 \cdot 10^{-11}$
700	$4.785 \cdot 10^{-7}$	$2.982 \cdot 10^{-8}$	$1.858 \cdot 10^{-9}$	$1.158 \cdot 10^{-10}$
800	$5.483 \cdot 10^{-7}$	$3.418 \cdot 10^{-8}$	$2.131 \cdot 10^{-9}$	$1.328 \cdot 10^{-10}$
900	$6.182 \cdot 10^{-7}$	$3.855 \cdot 10^{-8}$	$2.404 \cdot 10^{-9}$	$1.499 \cdot 10^{-10}$
1000	$6.880 \cdot 10^{-7}$	$4.291 \cdot 10^{-8}$	$2.677 \cdot 10^{-9}$	$1.669 \cdot 10^{-10}$

Table 5

Cardinalities of DNA chips composed of isothermic libraries of temperatures τ and $\tau + 2$

temperature τ	cardinality
26	1390592
28	3799168
30	10379520
32	28357376
34	77473792
36	211662336
38	578272256
40	1579869184

Table 6

Cardinalities of DNA chips composed of full library of oligonucleotides of length l

oligonucleotide length l	cardinality
8	65536
10	1048576
12	16777216
14	268435456
16	4294967296
18	68719476736
20	1099511627776
22	17592186044416

As one can notice the values of branching probability are higher for isothermic chips than for the classical ones of similar cardinality. The source of this phenomenon is the fact that the isothermic chip is composed of oligonucleotides of various lengths. There are oligonucleotides longer than those on the classical counterpart of the chip but there are also some shorter ones. Obviously, the probability of finding repetitions of these shorter oligonucleotides is higher than such a probability for oligonucleotides from the classical chip and it influences the overall branching probability of the isothermic chip.

5. Conclusions

In the paper a probabilistic analysis of the resolving power of isothermic DNA sequencing chips has been presented. These chips have been designed in order to reduce the ratio of hybridization errors. The analysis of chips' resolving power is based on the probability of appearing negative errors following from repetitions when the target DNA is sequenced by a given chip, which is called the branching probability. This probability can be a measure of the usefulness of the chip. The lower the value of branching probability the better the chip is suited for DNA sequencing. The values of the probability decreases when the lengths of the oligonucleotides on the chip increases. Unfortunately, currently it is impossible to construct classical chips composed of full libraries of oligonucleotides of lengths exceeding 10 nucleotides, which follows from technological constraints. Isothermic chips are composed of libraries consisting oligonucleotides of equal melting temperature, not length like in the classical case. The oligonucleotide composition of such libraries leads to a reduction of a number of positive and negative hybridization errors which is the most important feature of the libraries. On the other hand, the analysis of branching probability values for various cardinalities of classical and isothermic chips has shown that the probability of negative errors followed from subsequence repetitions is greater in the case of the isothermic chips. The result of the analysis may be a hint for those who would like to apply isothermic chips for DNA sequencing. The chips of this kind should be chosen carefully, since they do not reduce the numbers of all types of errors. In particular, they should not be applied for sequencing DNA molecules with many repetitions of short subsequences.

References

[1] R. Drmanac, I. Labat, I. Brukner and R. Crkvenjakov, "Sequencing of megabase plus DNA by hybridization: theory and the method", *Genomics* 4, 114–128 (1989).

[2] K. R. Khrapko, Y. P. Lysov, A. A. Khorlin, V. V. Shik, V. L. Florentiev and A. D. Mirzabekov, "An oligonucleotide approach to DNA sequencing", *FEBS Lett.* 256, 118–122 (1989).

[3] E. M. Southern, U. Maskos and J. K. Elder, "Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models", *Genomics* 13, 1008–1017 (1992).

[4] S. P. A. Fodor, J. L. Read, M. C. Pirrung, L. Stryer and A. Lu, D. Solas, "Light-directed, spatially addressable parallel chemical synthesis", *Science* 251, 767–773 (1991).

[5] E. M. Southern, United Kingdom patent application GB8 810400 (1988).

[6] A. C. Pease, D. Solas, E. Sullivan, M. Cronin, C. Holmes and S. Fodor, "Light-generated oligonucleotide arrays for rapid DNA sequence analysis", *Proc. Natl. Acad. Sci. USA* 91, 5022–5026 (1994).

[7] P. A. Pevzner and R. J. Lipshutz, "Towards DNA sequencing chips", in *Proceedings of 9th International Symposium MFCS'94*, ed. I. Prívara, B. Rovan, P. Ružička, pp. 143–151, August 22–26 (1994).

[8] J. D. Watson and F. H. C. Crick, "Genetic implications of the structure of deoxyribonucleic acid", *Nature* 171, 964–967 (1953).

[9] J. Błażewicz, P. Formanowicz, M. Kasprzak, W. T. Markiewicz and J. Węglarz, "DNA sequencing with positive and negative errors", *J. Comp. Biol.* 6, 113–123 (1999).

[10] J. Błażewicz, P. Formanowicz, M. Kasprzak and W. T. Markiewicz, "Sequencing by hybridization with isothermic oligonucleotide libraries", *Disc. Appl. Math.* 145, 40–51 (2004).

[11] R. B. Wallace, M. J. Johnson, T. Hirose, T. Miyake, E. H. Kawashima and K. Itakura, "The use of synthetic oligonucleotides as hybridization probes. ii. hybridization of oligonucleotides of mixed sequence to rabbit beta-globin DNA", *Nucleic Acids Res.* 9, 879–894 (1981).

[12] F. P. Preparata, A. M. Frieze and E. Upfal, "On the power of universal bases in sequencing by hybridization", in *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, ed. S. Istrail, P. Pevzner, M. Waterman, Lyon, France, pp. 295–301, April 11–14 (1999).

[13] P. Sachadyn and J. Kur, "Reducing the number of microlocations in oligonucleotide microchip matrices by the application of degenerate oligonucleotides", *J. Theor. Biol.* 197, 393–401 (1999).