Tabu search algorithm for DNA sequencing by hybridization with isothermic libraries

Jacek Blażejewicz‡, Piotr Formanowicz‡, Marta Kasprzak‡, Wojciech T. Markiewicz‡, Aleksandra Świercz‡

Abstract: In this paper, a problem of isothermic DNA sequencing by hybridization (SBH) is considered. In isothermic SBH a new type of oligonucleotide libraries is used. The library consists of oligonucleotides of different lengths depending on an oligonucleotide content. It is assumed that every oligonucleotide in such a library has an equal melting temperature. Each nucleotide adds its increment to the oligonucleotide temperature and it is assumed that A and T add 2°C and C and G add 4°C. The hybridization experiment using isothermic libraries should provide data with a lower number of errors due to an expected similarity of melting temperatures. From the computational point of view the problem of isothermic DNA sequencing with errors is hard, similarly like its classical counterpart. Hence, there is a need for developing heuristic algorithms that construct good suboptimal solutions. The aim of the paper is to propose a heuristic algorithm based on tabu search approach. The algorithm solves the problem with both positive and negative errors. Results of an extensive computational experiment are presented, which prove the high quality of the proposed method.

Keywords: sequencing by hybridization, isothermic libraries, heuristics, tabu search.

1 Introduction

One of the most important issues in molecular biology and genetics is reading DNA sequences. Despite the impressive achievements in sequencing whole genomes, there is still a need for an improvement of the existing sequencing methods or for a development of the new ones. One of the sequencing method is sequencing by hybridization proposed, among others, by Drmanac et al. [9], Khrapko et al. [18] and Southern et al. [27]. The method consists of two phases: the biochemical one and the computational one. In the biochemical stage, hybridization experiment is performed, in which an oligonucleotide library is compared with many copies of one strand of the examined DNA molecule. Usually, the library consists of all 4^l oligonucleotides of length l. The library is prepared on a DNA chip [26, 10, 21]. During the hybridization reaction fragments of the target DNA join to oligonucleotides from the library in their complementary places. After the reaction, one obtains the set of sequences being subfragments of the examined DNA molecule by reading a fluorescent or radioactive image of the chip. This set is named a spectrum. Most of the sequencing methods based on the hybridization reaction use sets of oligonucleotides of the same length. (An interesting exception are so called gapped chips [23].) Next the computational stage of the SBH comes.

It is known that in the case of an ideal hybridization experiment (no experimental errors involved), an original DNA sequence may be reconstructed in polynomial time [22]. On the other hand, if there are errors in the hybridization experiment (positive - excess of l-mers or negative - lack of l-mers), the computational phase of the method becomes a strongly NP-hard problem [14, 8]. The computational phase of the SBH approach with some restricted models of errors was addressed in a few papers [2, 18, 9, 22, 1, 16, 19, 7, 11, 12], and recently in [24, 25, 17]. The only exact method allowing for any type of error (positive and negative) and requiring no additional information about the spectrum was presented in [5]. Although this method is very expensive, an excessive number of errors in the spectrum makes this approach rather slow. Hence, a challenging problem is to design a new approach reducing a number of errors in the biochemical phase of SBH. Such an approach has been proposed in [3, 4]. It

---

*This material is based upon work supported by the National Science Foundation under Agreement No. 0112050.
†Institute of Computing Science, Poznań University of Technology, Piotrowo 3A, 60-965 Poznań, Poland.
‡Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowski 12/14, 61-704 Poznań, Poland.
§Corresponding author (Institute of Computing Science, Poznań University of Technology, Piotrowo 3A, 60-965 Poznań, Poland; e-mail: piotr@cs.put.poznan.pl).
is based on the usage of chip libraries consisting of oligonucleotides of equal melting temperatures. The approach follows a reasoning briefly described below. (More details can be found in [4].)

It is known that duplexes of C/G rich oligonucleotides are more stable than A/T rich ones. This phenomenon may result in numerous errors of the positive type. At the same time a modification of hybridization conditions directed towards diminishing the number of imperfect duplexes of C/G rich oligonucleotides may result in increasing a number of missing perfect duplexes of A/T rich oligonucleotides. This means that the modification decreasing a number of positive experimental errors might increase a number of negative ones. The duplex formation depends on a base composition but also on its length. In some models a simple equation was used to calculate melting temperatures of oligonucleotide duplexes assuming 4 degrees for C/G pairs and 2 degrees for A/T pairs [28]. Thus, at least formally, it is possible to compensate lower stability of A/T rich duplexes by increasing their length. It is known, that the above description is not very accurate although it reflects a general relative stability of different duplexes quite well. Therefore, to form more stable duplexes with hybridized sequences, in [3, 4] a new approach to SBH has been proposed. The key idea is to obtain a set of oligonucleotides that differ in base composition and length, and are characterized by a predefined relation between base composition and a length of oligonucleotides. In a specific case, if in a library an increment of C or G is twice of A or T and the sum of increments for each oligonucleotide is constant, then such a library is called isothermic. It may be proved that two such libraries are sufficient to reconstruct DNA sequences by the SBH method [4].

The computational phase of the reconstruction of the target sequence is strongly NP-hard if there are errors in the data [4] (similarly, like in the classical case). Hence, there is a need for developing efficient heuristics. The aim of this paper is to present such a method based on tabu search approach. The presented algorithm has been tested in an extensive computational experiment which confirmed a high quality of solutions as compared with those obtained by other methods using standard oligonucleotide libraries.

The organization of the paper is as follows. In Section 2 the problem is formulated and basic properties of isothermic libraries are recalled. In Section 3 the algorithm is presented, while in Section 4 the results of the computational experiment are shown. The paper ends with conclusions in Section 5.

2 Problem formulation

In this section we recall basic properties of isothermic libraries. More detailed analysis of their design and feature can be found in [4].

Let us start with a formal definition of isothermic oligonucleotide libraries [3, 4]. In this definition $w_i$ denotes an increment a nucleotide of type $i$ ($i \in \{A, C, G, T\}$) adds to the melting temperature of the oligonucleotide containing it. On the other hand, $x_i$ denotes a number of appearances of a nucleotide of type $i$ in the oligonucleotide.

**Definition**

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

Without loss of generality we may assume that $w_A = w_T = 2$ and $w_C = w_G = 4$. These increments correspond to temperatures which nucleotides bring into the stability of oligonucleotide duplexes. In what follows, a sum of increments of nucleotides forming an oligonucleotide will be called the oligonucleotide temperature.

Isothermic oligonucleotide library follows the experimentally established relationship between base composition and duplex stability described in Introduction. Oligonucleotides contained in such a library should form duplexes with their complements in a more narrow range of experimental conditions (temperature, salt concentration etc.) than that characteristic for an oligonucleotide library with oligomers of the same length. Therefore, the hybridization experiments performed with isothermic libraries should result in a smaller number of experimental errors. The use of such libraries should substantially limit a number of these errors to be considered in the computational phase of the SBH approach.
The basic question arising here is whether the isothermic library defined above is really well suited for SBH? One can easily realize that one such a library may be not enough. For example, it is not possible to cover DNA chains composed of only C and G nucleotides by oligomers from a library of a temperature not divisible by 4. Moreover, a library of a temperature divisible by 4 is not sufficient to cover sequences where one nucleotide of A or T type is surrounded by only G or C nucleotides. But, on the other hand, two such libraries with temperatures differing by one increment of A or T, i.e. by 2 deg, are sufficient to perform a proper hybridization experiment. More formally, the following theorem was proved (the proof can be found in [4]):

**Theorem 1** It is always possible to cover any DNA sequence by probes coming from two isothermic libraries of temperatures differing by 2 deg. Moreover, this coverage is such that in the sequence two consecutive oligonucleotides (from the libraries) have starting points shifted by at most one position.

The cardinality of the isothermic library of temperature \( t \) is between the cardinalities of the standard libraries involving, respectively, only the shortest \( \left( \frac{t}{4} \right) \) or only the longest \( \left( \frac{t}{2} \right) \) oligomers of equal length.

Now, we will formulate isothermic sequencing problem in its most general form, i.e. with positive and negative errors. The formulation will be valid under a reasonable assumption that most of the data coming from the hybridization experiment are correct. As in the classical SBH, as a result of a biochemical phase one gets a set of oligonucleotides that hybridized with the reconstructed DNA chain, i.e. spectrum \( (S) \). Now, the spectrum contains data from the hybridization experiment with two isothermic oligonucleotide libraries differing by one increment of A(T) nucleotide. Spectrum may contain two types of errors: positive, i.e. some oligonucleotides which are not a part of the DNA sequence appear, and negative, i.e. a lack of some oligonucleotides being a part of this sequence (including subsequent repetitions of the oligonucleotide). The most general problem is the one where both types of errors occur. This problem in a search version may be viewed as the one of finding, for a given spectrum \( S \), a sequence with the minimum number of positive and negative errors. It is not hard to see, that in case of the standard oligonucleotide library, where all oligonucleotides are of equal length, this formulation is equivalent to a maximization of a number of \( l \)-mers from the spectrum used to build a solution (a reconstructed sequence).

Let us denote by \( \alpha \) a number of oligonucleotides from the spectrum being a part of the constructed sequence. On the other hand, let \( \beta \) denote a number of oligonucleotides being members of the two used isothermic libraries (of temperatures respectively equal to \( t \) and \( t+2 \)), which can be distinguished in this sequence (each oligonucleotide adds to \( \beta \) a number of its occurrences in the sequence). We see that numbers of negative and positive errors in the constructed sequence may be defined as \( \beta - \alpha \) and \( |S| - \alpha \), respectively.

It is assumed here that the only information provided by the hybridization experiment are a spectrum and length \( n \) of the DNA sequence which is looked for. Having all the discussed notions in mind we can define our problem as follows.

**Isothermic DNA sequencing with negative and positive errors – search version.**

**Instance:** set \( S \) (spectrum) of oligonucleotides, each of them of temperature \( t \) or \( t+2 \), length \( n \) of an original sequence.

**Answer:** a sequence of length \( n \) with a minimum value of \( \beta + |S| - 2\alpha \).

The above problem may be proved to be strongly NP-hard [4], thus, the need arises to construct efficient algorithms constructing solutions (close to optimum) in a reasonable time. One of such method is presented in the next section.

### 3 The method

The heuristic algorithm for isothermic DNA sequencing presented in this section is based on tabu search metaheuristic being one of the most frequently used in combinatorial optimization [15]. In the
algorithm, spectrum consists of two data structures: an ordered list of oligonucleotides representing a current solution, and an unordered set of remaining oligonucleotides, called a trash. At each stage of the computations the number of oligonucleotides being on the list cannot be greater than the one that would produce a DNA sequence composed of at most \( n \) nucleotides (assuming that neighbors on the list are maximally overlapped). Moves obeying this constraint are called feasible and only such moves are performed by the method. The algorithm starts with the list constructed by a greedy selection of oligonucleotides, overlapping their neighbors.

In the method, three kinds of moves are initially defined: an insertion - a move transferring an oligonucleotide from the trash into the list, a deletion - a move transferring an oligonucleotide from the list to the trash, and a shift - a move transferring an oligonucleotide from one position of the list into another one.

Since moves performed on single oligonucleotides are usually not sufficient for generating good solutions, moves executed on clusters have been introduced. A cluster is a group of neighboring oligonucleotides in the solution (i.e. on the list) linked together. An overlap of each pair of neighboring oligonucleotides in the cluster is equal to \( p \) positions, where \( p \geq l(s_r) - 1 \), and \( l(s_r) \) is a length of the right oligonucleotide in the pair. Since insertion, deletion and shift may change a composition of a cluster, a list of clusters is updated after every move. Clusters exist only in the solution; they are broken into separate oligonucleotides once transferred into the trash. Finally, the set of moves has been defined as follows: insertion of an oligonucleotide, deletion of an oligonucleotide or a cluster, shift of an oligonucleotide or a cluster.

The following rules, limiting an application of the moves, have been introduced:
- a cluster may be shifted only if it does not break another cluster,
- an oligonucleotide may be shifted only if it is not a part of a cluster, provided it does not break any cluster,
- an oligonucleotide may be deleted only if it is located outside a cluster or at one of the ends of some cluster.

The algorithm tends to maximize a number of oligonucleotides in the solution. This maximization may be viewed as a global criterion function for the method, easier to implement in the context of tabu search and also leading to a reconstruction of the desired sequence. On the other hand, the only function which is able to compare all kinds of moves is a condensation defined for each solution to be the ratio of the number of oligonucleotides in the solution (taken from the spectrum) to the number of nucleotides in the solution (being a length of a currently constructed sequence). If the moves were compared by the global criterion function, deletion and shifts would be used very rarely. The maximization of the condensation causes a transformation of an initial solution into a series of collections of well-matched oligonucleotides. Note, that using the condensation as the only criterion for a choice of a move, would lead after a number of iterations to a creation of one cluster of length (in nucleotides) much less than \( n \). To avoid this fact, two approaches are used which cause a gradual increase of the number of oligonucleotides in the solution:

1° If a maximal value of the condensation is achieved by more than one move, the algorithm selects a move resulting in a greater number of oligonucleotides in the solution. Hence, the insertion is the most preferred move with shifts, the deletion of an oligonucleotide and the deletion of a cluster being next.

2° After a given number of moves without any improvement of the global criterion function value, the method executes a selected number of extending moves. Extending moves are feasible moves selected by using frequency-based memory instead of the condensation function. This memory is a structure that remembers the number of times each oligonucleotide from the spectrum appears in solutions. (For example, an oligonucleotide contained in all solutions generated so far has a frequency value equal to the number of iterations of the algorithm, while an oligonucleotide never used for constructing solutions has a frequency value equal to 0.) There are two types of extending moves: the insertion and the deletion of an oligonucleotide. The most preferred one is the insertion, and an oligonucleotide with the lowest frequency value is chosen. If no insertion is possible, an oligonucleotide with the highest frequency value is deleted from the solution. After the execution of extending moves, the method returns to the normal scheme with the condensation as a criterion function.

Such a combination of condensing and extending the solution, guarantees that the number of oligonucleotides will increase from a small value in the initial solution to a near-optimal value in the final one. The use of the frequency-based memory forces an inclusion of oligonucleotides that are not
well-matched into a solution. This effect would never be obtained by using the condensation function only.

Inserted or shifted oligonucleotides are remembered on the *tabu list* for a given number of iterations. The list is checked if an attempt to shift or to delete an oligonucleotide is made. The shift or deletion is prevented if the oligonucleotide is found to be on the list. Among others, this procedure makes it impossible to reverse extending moves. It does not appear necessary to remember clusters shifted, since clusters are often changed after a move. Even if the method gets into a cycle, it will stay there only briefly because of the next series of extending moves. An element on the tabu list may be deleted or shifted together with the cluster containing it. The element may also be deleted if there is no other feasible move. In such a case, an element that has been on the tabu list during the greatest number of iterations, is chosen. On the other hand, a deletion of a recently inserted oligonucleotide is not profitable from the viewpoint of a maximization of the global criterion function. Thus, an *aspiration criterion* for overriding the tabu status of an element is not used here.

To increase the quality of a solution, the algorithm does not stop after a given number of cycles consisting of *condensing* and *extending moves*, but restarts the search process. This is done by introducing *external cycles*. A new starting solution is generated on the basis of the frequency values, choosing oligonucleotides having the lowest frequencies. After that most of algorithm variables are set to initial values, thus, the search process is independent of the previous stages.

The first oligonucleotide of the target sequence can be an optional parameter for the method. The assumption concerning *its knowledge* is well justified since the *polymerase chain reaction* (PCR) is presently most often used to amplify a DNA to be sequenced. The tabu search method does not need to know the first oligonucleotide, but the lack of this knowledge may cause the algorithm to execute more iterations.

The method stops after a given number of the external cycles and returns as its outcome the sequence that has been found to contain the greatest number of oligonucleotides from the spectrum. A pseudocode of the described algorithm is given below.

**tabu search algorithm**

```plaintext
var: x, y, z, best, number: integer;
begin
    start from an initial solution, a permutation of oligonucleotides, that is obtained
    after applying greedy heuristic
    number := 0; best := 0;
    for y = 1 to a number of restarts
        begin
            for x = 1 to a number of condensing + extending cycles
                begin
                    iter := 0;
                    while iter != a number of condensing moves
                        begin
                            iter := iter + 1;
                            distinguish clusters in the solution;
                            execute the best feasible move (with the highest value of Condensation function);
                            number := a number of oligonucleotides in the solution;
                            when shifting or inserting an oligonucleotide put the oligonucleotide into the tabu
                            list for E iterations;
                            if best < number then
                                save the solution;
                                best := number;
                                iter := 0;
                            end
                        end;
            for z = 1 to a number of extending moves
                begin
                    insert an oligonucleotide which has been the most rarely used to build
```

5
the solution;
when inserting an oligonucleotide
    put the oligonucleotide into the tabu list for $E$ iterations;
    $number := number + 1$;
if no insertion is possible then
    delete the oligonucleotide that has been the most frequently used in the solution;
    $number := number - 1$;
if $best < number$ then
    begin
        $best := number$;
        save the solution;
    end;
$z := z + 1$
end;
// create new initial solution
put all oligonucleotides back from the solution to the trash set
(solution is an empty sequence);
repeat
    move the oligonucleotide from the trash set, which has been the most rarely used in
    the solution, at the end of the solution;
    if the sequence built so far is longer than $n$ then
        begin
            move the last oligonucleotide from the solution back to the trash set;
            break;
        end;
    until the sequence is longer than $n$ or all oligonucleotides are used
    almost all variables are set to initial values;
$y := y + 1$;
end;
end;

4 Computational experiments

The presented algorithm has been tested in an extensive computational experiment performed on PC
Pentium 4 machine with 256 kB RAM and Linux operating system. 10 DNA sequences have been
taken from GenBank as a basis for the testing. Their accession numbers are given in the Appendix.

From each of the sequences we created three sequences of length 200, 400, and 600 nucleotides,
respectively. These sequences were prefixes of the ones taken from GenBank. Next, a hybridization
experiment has been simulated by cutting off (from these sequences) oligonucleotides of melting tem-
peratures 26 and 28 deg, and, respectively, 36 and 38 deg. In this way two spectra for every sequence
were created, containing oligonucleotides of these temperatures.

The following idea has been used to introduce errors into the spectrum. The notation “200-
10%+10%” means that from an ideal spectrum (without any errors) for a sequence of length 200,
10% of randomly selected (according to a uniform distribution) oligonucleotides (negative errors) have
been deleted, and next 10% randomly generated oligonucleotides (positive errors) have been inserted.
Moreover, in all the cases a starting oligonucleotide could not be deleted and oligonucleotides which
have been introduced as positive errors had to be different from those already existing in the spectrum.
Errors have been generated in a progressive manner. The set “200-20%+20%” has the same errors as
the set “200-10%+10%”, but in addition 10% of negative errors and 10% of positive errors have been
inserted into it.

The algorithm requires seven parameters:
- a length of an original sequence $n$;
- a number of condensing moves performed without improving the value of the global criterion function (A);
- a number of extending moves (B);
- a number of cycles of the type “condensing moves + extending moves” (C);
- a number of external cycles (D);
- a length of the tabu list (E);
- (optionally) the first oligonucleotide in an original sequence.

In the experiments with oligonucleotides of temperatures 26 and 28 °C the following values of the parameters of the tabu method have been used: A = 7, B = 8, C = 600, D = 4 and E = 30. In the case of libraries of temperatures 36 and 38 °C the parameters were: A = 10, B = 7, C = 400, D = 4 and E = 30. These values were chosen after a series of preliminary tests.

It has been assumed that the first oligonucleotide of the reconstructed sequence is known. (As we already said, this assumption can be easily omitted at the expense of the time used by the algorithm.) The first (starting) solution has been generated by a simple heuristic algorithm [5].

A validation of the sequences generated by the method is done in two ways. Firstly, a usage of oligonucleotides from the spectrum (i.e. a percentage of a number of correct oligonucleotides from the spectrum composing a constructed sequence) is applied. Secondly, a similarity to the original sequence is checked using Needleman-Wunsch algorithm [20]. The algorithm compares two sequences: the one generated by the tabu search algorithm $s_t$ and the original sequence $s_o$. The similarity of the sequences is determined according to the following formula: $\sigma = 100\frac{2w - \psi}{\chi}$, where $\delta$ is a scoring for the two sequences, being a sum of scores for all columns in an optimal alignment (1 point for a match, -1 for mismatch or gap), and

$$
\psi = \begin{cases} 
    l(s_o)d + (l(s_t) - l(s_o))g & \text{if } l(s_t) > l(s_o) \\
    l(s_t)d + (l(s_o) - l(s_t))g & \text{otherwise}
\end{cases}
$$

$$
\chi = \begin{cases} 
    l(s_t)m & \text{if } l(s_t) > l(s_o) \\
    l(s_o)m & \text{otherwise}
\end{cases}
$$

where $l(s_t)$ and $l(s_o)$ are lengths of sequence $s_t$ and $s_o$, respectively, and $m = 1, d = -1, g = -1$.

The presented algorithm has been compared with the results of other competitive approaches for SBH with standard equal-length libraries. In fact, only few of the existing algorithms deal with both types of errors and non-constrained error characteristic [5, 6, 29]. Below such a comparison has been made between the presented algorithm and the one proposed in [6]. The latter is based on the formulation given in [5] reducing the SBH problem (with standard libraries) to a variant of the Selective Traveling Salesman Problem.

Table 1 contains the results of the computations done with the spectra of temperatures 26 and 28 deg. Each entry in the table is computed on the basis of the results obtained for sequencing the 10 DNA chains. For each chain and entry the sequencing process has been executed 100 times (for the same set of errors) to obtain a representative set of data. Hence, each entry is the average obtained from 1000 executions. These instances have been generated in order to compare results of our algorithm with the older one [6], basing on spectra of constant oligonucleotide length. (The reason for this is that two libraries of temperatures 26 and 28 deg have a similar cardinality as one library of the length equal to 10 (assumed in the experiment from [6]).) The results from [6] are presented in Table 2.

The computational experiments have been done on different machines, however, we can see that the algorithm from [6] consumes much more time than the new one with the growth of an instance size. The quality of its solutions is also much lower than in the case of the new algorithm. Because the main scheme of both algorithms is very similar, we could pose that the new isothermic approach has an advantage over the traditional one from the computational point of view.

The comparison of our results with the ones presented in [29] is possible only in part. Concerning random negative errors, they introduced only 5% of them, therefore only the instances 200-5%+5% and 400-5%+5% are comparable. Their results have been obtained in a shorter computation time. The quality of the solutions, however, cannot be compared since they did not give any such measure.

The approach presented in the paper works better if the temperatures assumed are higher. (Cardinalities of libraries are also bigger.) The results for the libraries of temperatures 36 and 38 deg are given in Table 3.

The results produced by the algorithm (Table 3) are of very high similarity to the original sequences.
<table>
<thead>
<tr>
<th>Instances</th>
<th>Usage [%]</th>
<th>Similarity [%]</th>
<th>Length [nucl.]</th>
<th>Time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>200-5%+5%</td>
<td>98.23</td>
<td>99.90</td>
<td>199.80</td>
<td>10.13</td>
</tr>
<tr>
<td>200-10%+10%</td>
<td>98.75</td>
<td>99.90</td>
<td>199.80</td>
<td>17.37</td>
</tr>
<tr>
<td>200-20%+20%</td>
<td>92.11</td>
<td>93.25</td>
<td>199.50</td>
<td>43.10</td>
</tr>
<tr>
<td>400-5%+5%</td>
<td>96.65</td>
<td>92.58</td>
<td>399.40</td>
<td>57.74</td>
</tr>
<tr>
<td>400-10%+10%</td>
<td>97.84</td>
<td>81.82</td>
<td>399.20</td>
<td>94.21</td>
</tr>
<tr>
<td>400-20%+20%</td>
<td>90.66</td>
<td>73.76</td>
<td>399.30</td>
<td>219.67</td>
</tr>
<tr>
<td>600-5%+5%</td>
<td>95.75</td>
<td>81.88</td>
<td>598.80</td>
<td>115.94</td>
</tr>
<tr>
<td>600-10%+10%</td>
<td>97.30</td>
<td>80.31</td>
<td>596.30</td>
<td>223.19</td>
</tr>
<tr>
<td>600-20%+20%</td>
<td>89.41</td>
<td>76.47</td>
<td>598.70</td>
<td>483.21</td>
</tr>
</tbody>
</table>

Table 1: Results for the spectra of temperatures 26 and 28 deg

<table>
<thead>
<tr>
<th>Instances</th>
<th>Usage [%]</th>
<th>Time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>200-5%+5%</td>
<td>94.63</td>
<td>54.2</td>
</tr>
<tr>
<td>200-10%+10%</td>
<td>93.50</td>
<td>66.0</td>
</tr>
<tr>
<td>200-20%+20%</td>
<td>92.00</td>
<td>86.4</td>
</tr>
<tr>
<td>400-5%+5%</td>
<td>87.92</td>
<td>790.2</td>
</tr>
<tr>
<td>400-10%+10%</td>
<td>89.64</td>
<td>830.7</td>
</tr>
</tbody>
</table>

Table 2: Results from [6] for the spectra of constant oligonucleotide length equal 10

<table>
<thead>
<tr>
<th>Instances</th>
<th>Usage [%]</th>
<th>Similarity [%]</th>
<th>Length [nucl.]</th>
<th>Time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>200-5%+5%</td>
<td>99.80</td>
<td>99.71</td>
<td>199.80</td>
<td>9.51</td>
</tr>
<tr>
<td>200-10%+10%</td>
<td>99.02</td>
<td>98.90</td>
<td>199.80</td>
<td>15.69</td>
</tr>
<tr>
<td>200-20%+20%</td>
<td>99.66</td>
<td>99.75</td>
<td>199.50</td>
<td>29.94</td>
</tr>
<tr>
<td>400-5%+5%</td>
<td>99.80</td>
<td>97.10</td>
<td>399.20</td>
<td>39.87</td>
</tr>
<tr>
<td>400-10%+10%</td>
<td>99.38</td>
<td>96.30</td>
<td>398.00</td>
<td>64.49</td>
</tr>
<tr>
<td>400-20%+20%</td>
<td>98.04</td>
<td>95.32</td>
<td>397.10</td>
<td>154.87</td>
</tr>
<tr>
<td>600-5%+5%</td>
<td>99.10</td>
<td>94.50</td>
<td>597.90</td>
<td>89.87</td>
</tr>
<tr>
<td>600-10%+10%</td>
<td>98.87</td>
<td>93.87</td>
<td>597.60</td>
<td>147.84</td>
</tr>
<tr>
<td>600-20%+20%</td>
<td>98.85</td>
<td>93.21</td>
<td>597.50</td>
<td>375.13</td>
</tr>
</tbody>
</table>

Table 3: Results for the spectra of temperatures 36 and 38 deg
and have been generated in relatively short (knowing requirements of the tabu search) computation times. It should be noted that the algorithm has a direct influence only on the usage of oligonucleotides from the spectrum and on the length of the solution (it maximizes the usage not exceeding the given length), and the values of the two parameters are very close to the optimal ones. The high values of the similarity to original sequences prove the correctness of the criterion functions implemented in the algorithm.

5 Conclusions

In the paper a heuristic algorithm for a new variant of DNA sequencing by hybridization has been proposed. The variant is based on libraries composed of oligonucleotides of equal melting temperatures (instead of equal length like in the classical approach). The use of such libraries may lead to a reduction of experimental error rate. Another advantage of the new approach follows from the observation that the libraries of the new type are composed of oligonucleotides of various lengths. Moreover, the cardinality of such a library is between the cardinality of a classic library composed of oligonucleotides of length corresponding to the shortest element from the new one and a library composed of elements of length corresponding to the longest element. Since the new library contains short and long oligonucleotides, the possibility of wrong reconstruction caused by repetitions decreases.

The computational phase of the new variant of the sequencing method is a computationally hard problem if there are errors in the input data (similarly to its classical counterpart). Hence, the need of developing good polynomial time heuristics arises. In the paper, an algorithm based on tabu search approach has been presented and its effectiveness has been tested in an extensive computational experiment. The results obtained in the experiment confirmed a high quality of the solutions provided by the algorithm, as compared with other existing algorithms for standard SBH.

References


Appendix

Accession numbers of sequences coding human proteins used in the computational experiments.

NM_005273, NM_016183, NM_080601, X51535, X56088, X03350, Y00264, Y00649, X05299, AF435957.