

## Algorithm

# Dealing with repetitions in sequencing by hybridization

Jacek Blazewicz<sup>a,b</sup>, Fred Glover<sup>c</sup>, Marta Kasprzak<sup>a,b</sup>, Wojciech T. Markiewicz<sup>b</sup>, Ceyda Oğuz<sup>d</sup>,  
Dietrich Rebholz-Schuhmann<sup>e</sup>, Aleksandra Swiercz<sup>a,b,\*</sup>

<sup>a</sup> Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland

<sup>b</sup> Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12, 61-704 Poznań, Poland

<sup>c</sup> University of Colorado, Boulder, CO 80309-0419, USA

<sup>d</sup> Department of Industrial Engineering, Koç University, Rumeli Feneri Yolu, 34450 Sariyer, Istanbul, Turkey

<sup>e</sup> European Bioinformatics Institute, Cambridge, UK

Received 3 February 2006; received in revised form 24 May 2006; accepted 24 May 2006

## Abstract

DNA sequencing by hybridization (SBH) induces errors in the biochemical experiment. Some of them are random and disappear when the experiment is repeated. Others are systematic, involving repetitions in the probes of the target sequence. A good method for solving SBH problems must deal with both types of errors. In this work we propose a new hybrid genetic algorithm for isothermic and standard sequencing that incorporates the concept of structured combinations. The algorithm is then compared with other methods designed for handling errors that arise in standard and isothermic SBH approaches. DNA sequences used for testing are taken from GenBank. The set of instances for testing was divided into two groups. The first group consisted of sequences containing positive and negative errors in the spectrum, at a rate of up to 20%, excluding errors coming from repetitions. The second group consisted of sequences containing repeated oligonucleotides, and containing additional errors up to 5% added into the spectra. Our new method outperforms the best alternative procedures for both data sets. Moreover, the method produces solutions exhibiting extremely high degree of similarity to the target sequences in the cases without repetitions, which is an important outcome for biologists. The spectra prepared from the sequences taken from GenBank are available on our website <http://bio.cs.put.poznan.pl/>.  
© 2006 Elsevier Ltd. All rights reserved.

**Keywords:** Sequencing by hybridization; Oligonucleotide libraries; Repetitions; Genetic algorithm

## 1. Introduction

DNA sequencing by hybridization (SBH) was proposed by Southern (1988), Bains and Smith (1988) and Drmanac et al. (1989) as an alternative to gel-based methods (Maxam and Gilbert, 1977; Sanger et al., 1977). Although it is not yet widely used because of current difficulties in performing biochemical experiments with the procedure, SBH is expected to be used more and more in the future (Fogel et al., 1998; Blazewicz et al., 1999b; Shamir and Tsur, 2002; Halperin et al., 2003; Fogel and Corne, 2003; Zhang et al., 2003; Heath et al., 2003; Blazewicz et al., 2004b). SBH consists of two steps. The first, biochemical one, is based on the principle of complementary hybridization

of two single-stranded DNA chains. A set of short, differing oligonucleotides (probes) is placed on a DNA chip. In the standard approach to DNA sequencing all the probes are strings of equal length  $l$ . Taking a full library of different oligonucleotides for the hybridization experiment, there are  $4^l$  different probes on the chip. These probes are compared with many copies of a fluorescently labeled DNA sequence (called the target sequence). During the biochemical experiment all oligonucleotides that are reverse complementary to the target sequence hybridize and emit a fluorescent signal. After the reaction, by reading the signals, one can obtain a set of probes that compose the DNA sequence. This set is called a spectrum. During the second step of the SBH, an algorithm reconstructs a sequence from the spectrum.

The ideal spectrum consists of all oligonucleotides that compose the target sequence and, in addition, either the spectrum contains no repeated probes or the multiplicity number of each  $l$ -mer is known. The problem of finding the target sequence can be reduced to the problem of finding certain combinatorial path in a labeled, directed graph (Lysov et al., 1988; Pevzner, 1989). A class of these graphs has recently been given the name DNA

\* Corresponding author. Tel.: +48 618528503; fax: +48 618771525.

E-mail addresses: [jblazewicz@cs.put.poznan.pl](mailto:jblazewicz@cs.put.poznan.pl) (J. Blazewicz), [Fred.Glover@Colorado.edu](mailto:Fred.Glover@Colorado.edu) (F. Glover), [mkasprzak@cs.put.poznan.pl](mailto:mkasprzak@cs.put.poznan.pl) (M. Kasprzak), [markiewicz@ibch.poznan.pl](mailto:markiewicz@ibch.poznan.pl) (W.T. Markiewicz), [coguz@ku.edu.tr](mailto:coguz@ku.edu.tr) (C. Oğuz), [rebholz@ebi.ac.uk](mailto:rebholz@ebi.ac.uk) (D. Rebholz-Schuhmann), [aswiercz@cs.put.poznan.pl](mailto:aswiercz@cs.put.poznan.pl) (A. Swiercz).

graphs and their properties have been analyzed by Blazewicz et al. (1999c).

However, in practice hybridization processes usually generate some errors. These errors can be of two types: *negative* errors, where some  $l$ -mers are missing in the spectrum, and *positive* errors, where some additional oligonucleotides appear in the spectrum. Negative errors can occur either as a result of wrong hybridizations or because some parts of the target sequence, at least as long as  $l$ , are repeated, to produce what are called *repetitions*.

We refer to the approach of sequencing by reference to probes of equal length  $l$  as the standard sequencing approach. The associated (standard) DNA sequencing problem has been proved to be strongly NP-hard in the case where errors occur in the spectrum (Blazewicz and Kasprzak, 2003). Several algorithms have been proposed for solving the sequencing problem, some of them restricted to problems with only one type of error (Bains and Smith, 1988; Pevzner, 1989; Drmanac et al., 1989; Guénoche, 1992; Blazewicz et al., 1997). The most general case was first studied in Blazewicz et al. (1999b), where a branch and bound algorithm was proposed, and subsequently in Zhang et al. (2003) and in Blazewicz et al. (2004b).

A few years ago new approaches to sequencing were proposed. Some of them are based on designing a new chip, which uses *universal* bases, i.e., nucleotides that hybridize with each of the standard bases (A, C, G or T) (Preparata et al., 1999a; Heath et al., 2003; Halperin et al., 2003). This approach uses gapped probing patterns instead of standard  $l$ -mers. One of the proposed gapped probes, GP( $s, r$ ), is of the form “ $X^s(N^{s-1}X)^r$ ”, where X is one of the nucleotides A, C, G, T, and N is the universal base. Thus, the number of specified nucleotides in the oligonucleotide is equal to  $k = s + r$ . If there are no restrictions on the set of probes, an information-theoretic argument yields an upper bound on the length of unambiguously reconstructible sequences:  $\theta(4^k)$  (Preparata et al., 1999b). In contrast, in a standard DNA sequencing problem the expected length of a sequence that can be unambiguously reconstructed with probes of length  $l$  is  $O(2^l)$ . Although the approach with gapped probes uses a small number of oligonucleotides in the hybridization experiment, it assumes the knowledge of the  $s(r + 1)$ -prefix and the procedure is vulnerable to error.

Another approach to SBH, called the *isothermic* approach, uses libraries of isothermic oligonucleotides, i.e., oligonucleotides of the same melting temperature, instead of those of an equal length  $l$ . It is well known that DNA duplexes of C/G rich  $l$ -mers are more stable than A/T rich  $l$ -mers. This obstacle can result in numerous errors in the spectrum. The dependence of the duplex formation on the base composition can be reduced by high concentration of chaotropic salt (Maskos and Southern, 1993). Moreover, by increasing the length of A/T rich duplexes, one can also increase their stability. In the earliest studies using allele specific oligonucleotides in DNA mutation analysis, a simple equation was used to count melting temperatures of oligonucleotide duplexes assuming  $4^\circ$  for every G/C pair and  $2^\circ$  for every A/T pair (Wallace et al., 1981). It is known that the above description is not very exact although it reflects a general relative stability of different duplexes quite well.

Oligonucleotides in an isothermic 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 oligonucleotides of the same length. Therefore, the hybridization experiments performed with isothermic libraries should result in a smaller number of experimental errors. Several computational methods were proposed for solving the isothermic sequencing problem (Blazewicz et al., 2004a, in press).

This paper first compares standard and isothermic approaches for dealing with the DNA sequencing problem with errors, especially errors coming from repetitions. While many methods exist for each approach, we restrict our attention to those that give the best results, as a basis for comparison. We cannot include the gapped-probe pattern approach in our study because prior studies either restrict attention to cases having a low error rate in data used for computations (Preparata et al., 1999a) or the results are presented using types of measures that cannot be compared with those customarily adopted (Preparata et al., 1999a; Halperin et al., 2003).

The computational tests designed for comparison were performed on the same target sequences for each method. Target sequences are DNA sequences that are coding human proteins and are taken from GenBank. For different approaches separate spectra were created with the same error rate, standard spectra with equal-length oligonucleotides, and isothermic spectra with oligonucleotides of the same temperature. The comparison of the selected methods (the best one for each approach) with such prepared spectra showed that isothermic spectra have more errors coming from repetitions than standard spectra, the reconstruction of an unknown sequence is then ambiguous and the probability of finding the correct sequence is smaller. We then propose a revised hybrid genetic algorithm, based on the main scheme of the hybrid genetic algorithm developed earlier for the isothermic approach (Blazewicz et al., in press). The results are improved significantly by this revised hybrid genetic algorithm using standard (i.e., equal length oligonucleotides) libraries, even for sequences with repeated oligonucleotides, where the average similarity to the target sequence is around 90% and almost half of the tested instances give the desired target sequence.

The next section defines isothermic libraries, describes the preparation of data for computational testing, and presents the results of comparison of the best algorithms for isothermic and standard sequencing. The following two sections describe our revised hybrid genetic algorithm and report the results of computational tests of the new method compared to other algorithms. The last section contains conclusions.

## 2. Comparison of standard and isothermic approaches to SBH

Among many approaches to SBH we focus on the alternative to the standard approach that makes use of isothermic libraries (Blazewicz et al., 1999a, 2004c). In the previous section we briefly described the isothermic oligonucleotide library. Now we give a more formal definition.

**Definition.** An *isothermic library*  $L$  of temperature  $t_L$  is a library of oligonucleotides satisfying relations

$$w_A x_A + w_C x_C + w_G x_G + w_T x_T = t_L,$$

$$w_A = w_T, \quad w_C = w_G \quad \text{and} \quad 2w_A = w_C,$$

where  $x_i$  is the number of occurrences of nucleotide of type  $i$  in the oligonucleotide, and  $w_i$  denotes an increment of a nucleotide of type  $i$  added to the melting temperature of an oligonucleotide ( $i \in \{A, C, G, T\}$ ).

Without loss of generality we assume that  $w_A = w_T = 2^\circ$  and  $w_C = w_G = 4^\circ$ . This corresponds to increments that allow nucleotides to form stable oligonucleotide duplexes. In what follows, the sum of increments of nucleotides forming an oligonucleotide will be called an oligonucleotide *temperature*.

In order to perform a proper hybridization experiment (when all oligonucleotides covering an analyzed DNA sequence can be detected) it is proved that it is sufficient to use two such libraries differing by one increment of A or T (by  $2^\circ$ ) and that one such library is not enough (Blazewicz et al., 2004c).

To evaluate the cardinality of isothermic libraries we use two equations—(1) for oligonucleotides with temperatures divisible by 4 and (2) for oligonucleotides with temperatures non-divisible by 4 (Blazewicz et al., 2004c).

$$\text{card}(t) = \sum_{i=0}^{t/4} \left[ \binom{t/4 + i}{2i} 2^{(t/4)+i} \right] \quad (1)$$

$$\text{card}(t) = \sum_{i=0}^{\lfloor t/4 \rfloor} \left[ \binom{\lfloor t/4 \rfloor + i + 1}{2i + 1} 2^{\lfloor t/4 \rfloor + i + 1} \right] \quad (2)$$

Now, the problem of isothermic SBH can be formulated as follows. As the input data one gives spectrum  $S$ , i.e., a set of oligonucleotides that possibly hybridize with the target sequence, and the length  $n$  of the sequence. The goal is to maximize the number of probes used to form the output sequence.

For comparison of the standard and isothermic approaches, we chose the best method for each approach. For standard libraries the one that achieves the best results is the tabu and scatter search algorithm that was described in Blazewicz et al. (2004b). The algorithm has the following structure. The spectrum consists of two sets: solution—an ordered list of oligonucleotides which compose a sequence not longer than  $n$ , and trash—an unordered set of remaining oligonucleotides. The oligonucleotides in the solution set that are well fitted to their neighbors (the  $l-1$  right nucleotides of an oligonucleotide are overlapping  $l-1$  left nucleotides of the next oligonucleotide) form a cluster, i.e., a block that cannot be broken during performing the next move. The clusters may change after each move. The algorithm starts with an initial solution and searches for the best solution in the neighborhood. After a chosen number of iterations where the number of oligonucleotides in the solution set does not increase, the method stops and re-starts the search in a different part of the solution space. During local search computations, a set of the best solutions, whose members are not too similar, is collected and serves as a reference set for creating a new starting point by the re-starting process. This algorithm achieved very good

results for tested instances, even with a high rate of errors in the spectrum, as it will be shown later.

The second algorithm, used for isothermic libraries, is the hybrid genetic algorithm presented in Blazewicz et al. (in press). This method operates on a population of individuals. Each individual is a permutation of all oligonucleotides from the spectrum. Fitness of an individual is the greatest number of neighboring oligonucleotides that form a sequence not longer than  $n$ . After selecting the individuals as parents, operators such as mutation and crossover are applied. The offspring inherits the best features of the parents, enhanced by the structured crossover; thus, the next generation is more adapted to the environment (i.e., the solution is composed of higher number of oligonucleotides). After a selected number of iterations without improvement, the algorithm stops and offers the best sequence found as its output. The algorithm often solves tested instances to optimality and the sequences obtained usually yield a 100% similarity measure by reference to the original sequences.

The tests of these two algorithms were divided into two stages based on two sets of instances. The first set, A, is generated from sequences with no repetitions of oligonucleotides. Additional errors were introduced into these spectra: 5% of negative errors and 5% of positive errors, or 20% and 20%, respectively. The second set, B, contains spectra with errors coming from repetitions. This set contains two different types of instances: one including only repetitions in the spectra and one containing additional positive and negative errors.

The preparation of *set A* proceeded as follows. Forty sequences coding human proteins were obtained from GenBank. (Their accession numbers can be found in Appendix A.) From prefixes of these sequences of length 200, 400, 500 and 600 nucleotides, respectively, two kinds of ideal spectra were created, one for a standard library with oligonucleotide length  $l = 10$  and one for an isothermic library with oligonucleotide temperatures  $t = 26^\circ$  and  $t + 2 = 28^\circ$ . According to Eqs. (1) and (2) the cardinality of this standard library is approximately the same as the sum of cardinalities of the two isothermic libraries. The spectra contained no repeated oligonucleotides. Next, random positive and negative errors were introduced into the spectra at the level of 5% or 20% of the initial cardinalities. Positive errors were compelled to be different from the oligonucleotides already present in the spectra.

*Set B* was prepared in a similar way. Forty sequences coding human proteins were obtained from GenBank, but this time their prefixes of length 600 nucleotides induced some repetitions both in isothermic and standard spectra. (Accession numbers of the sequences can be found in Appendix B.) Spectra created from these sequences contained some negative errors (repetitions). In order to choose a typical set of sequences for our experiment, we took the ones, which resulted in the same average number of repetitions as 1000 randomly chosen sequences coding human proteins from GenBank. For our instances generated with respect to the standard libraries the number of repetition errors varied from 1 to 17 probes where average for the spectrum was 4 repetitions. In the case of isothermic libraries the number of errors coming from repetitions for the spectra was 4–30 and the average was 16 oligonucleotides. Set B.1 contained only errors coming from

Table 1  
Results of tests for set A—sequences without repetitions

	Length of the sequence							
	200		400		500		600	
	5% <sup>a</sup>	20% <sup>a</sup>	5% <sup>a</sup>	20% <sup>a</sup>	5% <sup>a</sup>	20% <sup>a</sup>	5% <sup>a</sup>	20% <sup>a</sup>
Tabu and scatter search with standard libraries								
Usage of oligonucleotides (%)	99.93	99.93	99.90	99.67	99.83	99.64	99.84	99.36
Similarity (%)	99.87	98.44	98.96	95.70	95.76	92.41	95.82	88.50
No. of optimal solutions	39/40	38/40	37/40	28/40	32/40	27/40	32/40	19/40
Running time (s)	5.62	8.99	47.38	68.50	82.94	125.54	127.57	186.26
Hybrid genetic algorithm with isothermic libraries								
Usage of oligonucleotides (%)	100.00	100.00	100.00	99.99	100.00	99.98	100.00	99.98
Similarity (%)	99.94	99.20	99.21	99.18	99.81	99.59	97.96	97.97
No. of optimal solutions	39/40	37/40	38/40	36/40	39/40	35/40	36/40	32/40
Running time (s)	6.50	8.52	23.90	30.81	46.49	53.64	80.91	91.57

<sup>a</sup> Error rate.

repetitions while in set B.2 additional random negative errors up to 5% and additional 5% of positive errors were introduced.

For these spectra, the results of testing our two algorithms are presented in Tables 1 and 2. In the tables all entries, except the number of optimal solutions, are given as the average value of 40 different instances. The percentage of oligonucleotides from the spectrum used for composing the target sequence is presented in the row ‘usage of oligonucleotides’. The figure ‘100%’ means that the number of oligonucleotides from the spectrum is the same as the number of proper oligonucleotides, i.e., the cardinality of the ideal spectrum diminished by the number of negative errors. ‘Similarity’ is calculated according to the pairwise sequence alignment of Needleman–Wunsch algorithm (Needelman and Wunsch, 1970). The number of solved instances, where the sequence obtained has 100% similarity to the original one, is presented in the next row. ‘Running time’ is the average total time of computations made on Pentium 4, 2.0 GHz, with 512 MB RAM.

Analyzing the results, we can state that the hybrid genetic algorithm with isothermic libraries deals very well with the instances where no repetition of oligonucleotides appears, while in the case with repetitions the tabu and scatter search method with standard libraries works better. This might be for the following reason: isothermic libraries cause so many repetitions that solving the problem is much more difficult. In description of preparation of Set B (sequences with repetitions of oligonucleotides) it was mentioned that the average number of repetitions is 4 for standard libraries and 16 for isothermic libraries for the same sequences. Although isothermic libraries cause fewer experimental errors, they give more repetition errors than standard libraries, because an isothermic library consists of nucleotide probes shorter than 10 nucleotides for the GC rich sequences. As a result, one might suppose that connecting the standard libraries with hybrid genetic algorithm would give the best results. Although there were many different genetic algorithms applied to this problem (Blazewicz et al., 2002; Bui and Youssef, 2004) we propose a new one in the next section. The revised hybrid algorithm was based on the main pattern of the approach described in Blazewicz et al. (in press) but adopted to handle standard libraries.

### 3. Revised hybrid genetic algorithm

General idea of the genetic algorithm was proposed by Holland (1975), and has been applied to different combinatorial problems (Aarts and Lenstra, 1997; Voss et al., 1998). Specifically, the genetic algorithm is a mechanism that simulates natural evolutionary processes. Its basic components are: population, individuals (also called chromosomes), fitness of the individuals, reproduction process including selection of parents and generation of children (genetic operation), replacement and completion of generation processes. A typical genetic algorithm starts with an initial population of individuals representing possible solutions to the problem. Each individual is evaluated by its fitness, which is determined by the associated value of the objective function. The next generation (offspring) is created after applying genetic operators to the fittest individuals from the parent population. There are several different genetic operators. Among them there are crossover (adopting the common features of each parent and mixing the remaining features) and mutation (introducing variation into the individuals). Hence, new individuals will have somewhat different features compared to their parents. In a new generation, the fitness of the offspring is evaluated in a fashion similar to that for their parents. This birth process to-

Table 2  
Results of tests for set B.1—sequences of length 600 with repetitions of oligonucleotides without additional errors in spectra and for set B.2—sequences of length 600 with repetitions of oligonucleotides and additional up to 5% of negative and 5% of positive errors

Algorithm	Set B.1	Set B.2
Tabu and scatter search with standard libraries		
Usage of oligonucleotides (%)	99.86	99.55
Similarity (%)	88.45	82.63
No. of optimal solutions	14/40	10/40
Running time (s)	84.30	129.95
Hybrid genetic algorithm with isothermic libraries		
Usage of oligonucleotides (%)	99.99	100.00
Similarity (%)	78.34	79.12
No. of optimal solutions	2/40	3/40
Running time (s)	90.64	108.67

gether with a death process will define a generation, as well as a population size. However, the population size usually remains constant from one generation to the next. This procedure is repeated until a stopping criterion is reached. The output of the simulated evolution process of a genetic algorithm will be the best chromosome found, which can be a highly evolved solution to the problem. The effectiveness of the algorithm depends on how the particular components of genetic algorithm are designed. Below we describe the implementation of these issues in our proposed algorithm.

Our hybrid genetic algorithm adopts the general structure described except that we use a special type of crossover called structured crossover that shares ideas in common with those proposed in the setting of tabu search. We implemented each of the components of the algorithm by considering the characteristics of the DNA sequencing by hybridization problem as explained below:

*Data* coming from the hybridization experiment are the spectrum  $S$ , i.e., a set of oligonucleotides of the same length  $l$ , and the length of the original sequence  $n$ .

*Initial population* consists of  $s$  randomly generated individuals.  $s$  is the population size, and it is kept constant during computations.

*Individual (chromosome)* is a permutation of  $|S|$  indices of oligonucleotides from the spectrum. Every permutation is decoded into a sequence, usually longer than  $n$ , and its best subsequence of length not greater than  $n$ , i.e., including the largest number of oligonucleotides from the spectrum, is a potential solution.

*Fitness* (objective function) is the number of oligonucleotides from the spectrum used to form the best subsequence, which is not longer than  $n$ .

*Selection:* Individuals from the current population are selected based on their fitness values to form the mating pool for the reproduction step. The aim of the selection is to keep good individuals and to eliminate the bad ones from one generation to the other. This selection is performed by using the part sum selection procedure in our implementation. This procedure has some components of proportional selection schemes like roulette wheel selection and of remainder schemes like deterministic procedure. First, each individual in the population is evaluated according to its fitness value to obtain the probability of selecting this individual as a parent. Then, in a deterministic way, the individuals are selected.

*Reproduction:* New individuals (offspring) in the next generation are obtained by applying the operators: structured crossover and mutation to the mating pool obtained in the previous step, with the probability, respectively,  $c$  and  $m$ . Thus, in the next generation  $c \cdot s$  new individuals will be created with structured crossover and at most  $m \cdot s \cdot |S|$  new individuals will come from mutation. The probabilities ( $c$ ,  $m$ ) were determined during our preliminary tests to set their best combination. In the mutation operator, described below, we use the concept of overlap degree, which is the number of nucleotides overlapping in two ad-

acent oligonucleotides. We further define the total overlap degree of an oligonucleotide as the overlap degree with its predecessor plus the overlap degree with its successor.

*Structured crossover:* Two parents are chosen randomly from the mating pool. The first oligonucleotide in the offspring is chosen randomly. This oligonucleotide,  $o_i$ , is identified in both of the parents. The construction of the child from this starting point departs from that of classical crossover operations, and uses a strategic design that accords with the concept of *structured combinations* as introduced in Glover (1994). The construction proceeds as follows. The successors of  $o_i$  and the predecessors of  $o_i$  in the parent individuals are considered. The one that fits better in front of  $o_i$  (for all predecessors in the parent chromosome) or at the end of  $o_i$  (for all successors in the parent chromosome) is placed at the proper position in the offspring. The new oligonucleotide together with  $o_i$  in the offspring now form a *block*. In the next steps instead of  $o_i$ , the terminal oligonucleotides of the block are considered and predecessors of the first oligonucleotide of the block and successors of the last oligonucleotide of the block are checked in the parent individuals. If there is neither unused successor nor unused predecessor in the parent individuals, then the best fit oligonucleotide from the remaining individuals is chosen. The best fit oligonucleotide is the one with the lowest value of the ratio of the number of oligonucleotides in the block to the length of the sequence. This in fact means that the newly built sequence should be the shortest. The procedure of creating the individual is stopped when all oligonucleotides from the spectrum are in the block. In this formulation one child is created from two parents. As previously mentioned, the process of creating our new offspring is different from the classic genetic algorithm, where most of the process goes randomly. Our structured crossover operator incorporates the idea of structured combination by treating the individuals as *vectors* to establish precedence relationships between oligonucleotides. Selection of an oligonucleotide at each step can be compared to choosing the *best vote* from two vectors (parents). Hence, the offspring inherits the best characteristics (votes) from parents (vectors). Such an algorithmic construction process that uses “voting evaluations” based on the composition of the parents and the problem objective is also a basic feature of path relinking (Glover and Laguna, 1993).

*Mutation* can occur both in the parents and in the offspring population with the probability  $m$ . The individual to be mutated is selected randomly. In this individual the oligonucleotide with the lowest overlap degree is found (if more than one exist, the first one is chosen). This oligonucleotide is swapped with its neighbor that has lower overlap degree with it. If the selected oligonucleotide is the first (last) one, then it is swapped with the last (first) oligonucleotide. The new individual then replaces the old one.

*Creation of the next generation:* Apart from the newly created individuals, the best individuals from the parent population go on to the next generation.

*Stopping criterion* is a chosen number of generations in which no improvement occurs in the objective function.

After many preliminary tests a set of best parameter values was established. The cardinality of the population was set to the half of the length of the target sequence. For the shortest sequences  $s$  was equal to 100 individuals, and for the longest sequences 300 individuals. The probability that mutation of each oligonucleotide in the chromosome occurs was  $m = 0.001$ , and the probability that structured crossover occurs for any two individuals was  $c = 0.9$ , where the stopping criterion was set to be 50 generations without improvement of the objective function.

#### 4. Computational results of the revised hybrid genetic algorithm

Our revised hybrid genetic algorithm was tested with the same two sets of spectra, one without errors coming from repetitions (set A) and the other with errors coming from repetitions (set B) in spectra, that we used in our earlier comparisons of standard and isothermic approaches. The results for set A are presented in Table 3 and for set B in Table 4 as the average values calculated for 40 instances.

For comparison with our revised hybrid genetic algorithm for the sequences without repetition, among many algorithms dedicated for the problem of standard SBH we chose two algorithms. The first one is the combined tabu and scatter search method of Blazewicz et al. (2004b) as described earlier and the other is a genetic algorithm presented in Bui and Youssef (2004), referred to as the ‘enhanced genetic algorithm’. Comparison to some other algorithms was not possible because of different types of tested instances (Halperin et al., 2003) or because of different measures used to evaluate solutions (Zhang et al., 2003). The genetic algorithm, presented in Endo (2004), obtained the results (shown in brackets in his paper) that exceeded the maximal possible value.

The authors of the enhanced genetic algorithm (Bui and Youssef, 2004) introduced an additional step of preprocessing. They combine single oligonucleotides into ‘clusters’, in such a way that in one cluster neighboring oligonucleotides overlap with  $l - 1$  nucleotides. The idea of combining oligonucleotides into the cluster is very similar to the one described for tabu and scatter search, but in this enhanced genetic algorithm procedure, the clusters are fixed and the genetic operations are performed only on these clusters. After this initial step the classical genetic algorithm was applied with three-point crossover. The enhanced genetic algorithm was tested with a set of spectra coming from 40 sequences, different from ours, without repetitions of oligonucleotides. Spectra contained 20% of negative and 20% of positive errors; thus, comparison was not possible in every case.

Analyzing the results of tests with set A one can notice that our revised hybrid algorithm works very well even for very hard instances—for sequences of length 600 nucleotides and 20% of error rate. In almost all cases it finds the target sequence and almost all of the proper oligonucleotides were composing the solution. It improves significantly the results of tabu and scatter search and enhanced genetic algorithm for the problems in this test set, especially in the number of generated optimal solutions. Number of oligonucleotides from the spectrum used to form the solution is usually close to optimum. (Note that these results are also significantly better than the ones obtained for the previous version of this algorithm designed for the isothermic libraries, as given in the previous section.) The computation time of our algorithm was much shorter than for tabu and scatter search, especially for long sequences, but slightly longer than for enhanced genetic algorithm.

The results of tests performed on set B with our algorithm are presented in Table 4. Comparing these with the results of Table 2, we can notice that for each measure the results were improved, especially the number of optimal solutions increases to almost half of the tested instances. The similarities of obtained

Table 3  
Comparison of results of different algorithms for standard sequencing

	Length of the sequence							
	200		400		500		600	
	5% <sup>a</sup>	20% <sup>a</sup>	5% <sup>a</sup>	20% <sup>a</sup>	5% <sup>a</sup>	20% <sup>a</sup>	5% <sup>a</sup>	20% <sup>a</sup>
Revised hybrid genetic algorithm (PC Pentium 4, 2.0 GHz, 512 MB RAM)								
Usage of oligonucleotides (%)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.98
Similarity (%)	100.00	98.88	100.00	100.00	100.00	99.75	100.00	99.98
No. of optimal solutions	40/40	39/40	40/40	40/40	40/40	38/40	40/40	37/40
Running time (s)	4.30	6.32	11.90	20.46	18.07	27.19	25.69	44.19
Tabu and scatter search (Blazewicz et al., 2004b) (PC Pentium 4, 2.0 GHz, 512 MB RAM)								
Usage of oligonucleotides (%)	99.93	99.93	99.90	99.67	99.83	99.64	99.84	99.36
Similarity (%)	99.87	98.44	98.96	95.70	95.76	92.41	95.82	88.50
No. of optimal solutions	39/40	38/40	37/40	28/40	32/40	27/40	32/40	19/40
Running time (s)	5.62	8.99	47.38	68.50	82.94	125.54	127.57	186.26
Enhanced genetic algorithm (Bui and Youssef, 2004) (PC Pentium 4, 2.4 GHz, 512 MB RAM)								
Similarity (%)	–	97.60	–	92.90	–	92.00	–	–
No. of optimal solutions	–	26/40	–	13/40	–	13/40	–	–
Running time (s)	–	1.50	–	8.60	–	15.10	–	–

Sequences have no repetitions—set A.

<sup>a</sup> Error rate.

Table 4  
Results of tests of the revised hybrid genetic algorithm

	Set B.1	Set B.2
Usage of oligonucleotides	100.00	99.99
Similarity (%)	90.99	92.60
No. of optimal solutions	18/40	18/40
Running time	24.17	25.13

Set B.1—sequences with repetitions without additional errors in spectra. Set B.2—sequences with repetitions and additional negative and positive errors up to 5%.

sequences to the original ones were above 90%, which is very satisfying. We note that the earlier hybrid genetic algorithm designed for isothermic sequencing in Blazewicz et al. (in press) almost never found an original sequence for set B, although the main scheme is very similar to our algorithm. This confirms our earlier observation that isothermic libraries are more repetitive and produce more than one sequence with the same value of the objective function. Hence, the probability of finding a proper sequence is lower.

## 5. Conclusion

In this paper various methods and approaches to sequencing by hybridization were considered. The performance of the algorithms depended upon the nature of the sequences. In the instances where sequences have no repetitions the method that performs best is the hybrid genetic algorithm designed for isothermic sequencing (Blazewicz et al., in press). For problems that contain some repeated subsequences in the target sequence the tabu with scatter search method for standard sequencing (Blazewicz et al., 2004b) proved much better, finding optimal solutions with considerably greater frequency. This quite naturally suggested revising the hybrid genetic algorithm with a new crossover operator and combining with the standard library of oligonucleotides. This new revised hybrid genetic algorithm solves the SBH problem optimally in most of the cases (considering maximization of the objective function). In the case without repetitions the method found the same sequence as the target sequence, an outcome that is highly important for biologists. Moreover, for sequences having repetitions, it behaved extremely well, producing results outperforming by far all other existing algorithms.

## Acknowledgment

The research has been partially supported by the KBN grant no. 3T11F00227.

## Appendix A

Accession numbers of the sequences used for computational tests. Set A—sequences have no errors coming from repetitions: NM\_016055, NM\_016080, NM\_152373, NM\_002938, BC040844, NM\_152763, NG\_002692, BC044213, NG\_002660, NG\_002481, BC007770, BC015575, BC004538, BC063108, NG\_002361, NG\_001569, BC056270, NM\_153834, HSA519841, BC053904, BC062471, BC062325,

BC057825, NG\_001151, NG\_001292, NM\_005337, NM\_032293, NM\_032292, NM\_198197, NM\_032423, NM\_021807, NM\_015435, NM\_024622, NM\_030633, NM\_172366, NM\_177959, NM\_005337, NM\_003318, AF435957, AF497481.

## Appendix B

Accession numbers of the sequences used for computational tests. Set B—sequences contain errors coming from repetitions: NM\_002052, NM\_003008, NM\_183353, BC008923, BC012982, NG\_002363, BC009854, NM\_194310, NM\_006402, NM\_194300, NM\_020690, NM\_005745, BC000774, NM\_182498, BC047640, NM\_004902, NM\_020713, BC062620, NM\_032866, NG\_000980, BC026171, NM\_000321, BC063041, BC005805, NM\_144767, BC001077, NM\_032349, BC053852, NM\_021934, NM\_015698, BC050425, NM\_018046, NM\_004663, BC007398, NM\_016428, NM\_177423, BC026078, NM\_031205, BC041372, NM\_033318.

## References

- Aarts, E.H.L., Lenstra, J.K. (Eds.), 1997. Local Search in Combinatorial Optimization. John Wiley and Sons, Chichester.
- Bains, W., Smith, G.C., 1988. A novel method for nucleic acid sequence determination. *J. Theor. Biol.* 135, 303–307.
- Blazewicz, J., Formanowicz, P., Kasprzak, M., Markiewicz, W.T., 1999a. Method of sequencing of nucleic acids. Polish Patent Application P335786.
- Blazewicz, J., Formanowicz, P., Kasprzak, M., Markiewicz, W.T., Weglarz, J., 1999b. DNA sequencing with positive and negative errors. *J. Comput. Biol.* 6, 113–123.
- Blazewicz, J., Hertz, A., Kobler, D., de Werra, D., 1999c. On some properties of DNA graphs. *Discrete Appl. Math.* 98, 1–19.
- Blazewicz, J., Formanowicz, P., Kasprzak, M., Markiewicz, W.T., Swiercz, A., 2004a. Tabu search algorithm for DNA sequencing by hybridization with isothermic libraries. *Comput. Biol. Chem.* 28, 11–19.
- Blazewicz, J., Glover, F., Kasprzak, M., 2004b. DNA sequencing—tabu and scatter search combined. *INFORMS J. Comput.* 16, 232–240.
- Blazewicz, J., Formanowicz, P., Kasprzak, M., Markiewicz, W.T., 2004c. Sequencing by hybridization with isothermic oligonucleotide libraries. *Discrete Appl. Math.* 145, 40–51.
- Blazewicz, J., Kaczmarek, J., Kasprzak, M., Markiewicz, W.T., Weglarz, J., 1997. Sequential and parallel algorithms for DNA sequencing. *CABIOS* 13, 151–158.
- Blazewicz, J., Kasprzak, M., Kuroczycki, W., 2002. Hybrid genetic algorithm for DNA sequencing with errors. *J. Heuristics* 8, 495–502.
- Blazewicz, J., Kasprzak, M., 2003. Complexity of DNA sequencing by hybridization. *Theor. Comput. Sci.* 290, 1459–1473.
- Blazewicz, J., Oğuz, C., Swiercz, A., Weglarz, J. DNA sequencing by hybridization via genetic search. *Oper. Res.*, in press.
- Bui, T.N., Youssef, W.A., 2004. An enhanced genetic algorithm for DNA sequencing by hybridization with positive and negative errors. *Lect. Notes Comput. Sci.* 3103, 908–919.
- Drmanac, R., Labat, I., Brukner, I., Crkvenjakov, R., 1989. Sequencing of megabase plus DNA by hybridization: theory and method. *Genomics* 4, 114–128.
- Endo, T.A., 2004. Probabilistic nucleotide assembling method for sequencing by hybridisation. *Bioinformatics* 20, 2181–2188.
- Fogel, G.B., Chellapilla, K., Fogel, D.B., 1998. Reconstruction of DNA sequence information from a simulated DNA chip using evolutionary programming. *Proceedings of the 7th International Conference on Evolutionary Programming VII*, 429–436.

- Fogel, G.B., Corne, D.W. (Eds.), 2003. *Evolutionary Computations in Bioinformatics*. Morgan Kaufman, San Francisco.
- Glover, F., Laguna, M., 1993. Tabu search. In: Reeves, C.R. (Ed.), *Modern Heuristic Techniques for Combinatorial Problems*. Blackwell Scientific Publications, Oxford, pp. 70–150.
- Glover, F., 1994. Tabu search for nonlinear and parametric optimization (with links to genetic algorithm). *Discrete Appl. Math.* 49, 231–255.
- Guénoche, A., 1992. Can we recover a sequence, just knowing all its subsequences of given length? *CABIOS* 8, 569–574.
- Halperin, E., Halperin, S., Hartman, T., Shamir, R., 2003. Handling long targets and errors in sequencing by hybridization. *J. Comput. Biol.* 10, 483–497.
- Heath, S.A., Preparata, F.P., Young, J., 2003. Sequencing by hybridization by cooperating direct and reverse spectra. *J. Comput. Biol.* 10, 499–508.
- Holland, H., 1975. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor.
- Lysov, Y.P., Florentiev, V.L., Khorlin, A.A., Khrapko, K.R., Shik, V.V., Mirzabekov, A.D., 1988. Determination of the nucleotide sequence of DNA using hybridization of oligonucleotides. A new method. *Dokl. Akad. Nauk SSSR* 303, 1508–1511.
- Maskos, U., Southern, E.M., 1993. A study of oligonucleotide reassociation using large arrays of oligonucleotides synthesized on a glass support. *Nucleic Acids Res.* 21, 4663–4669.
- Maxam, A.M., Gilbert, W., 1977. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.* 74, 560–564.
- Needelman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities of the amino acid sequence of two proteins. *Proc. J. Mol. Biol.* 48, 443–453.
- Pevzner, P.A., 1989. *l*-Tuple DNA sequencing: computer analysis. *J. Biomol. Struct. Dyn.* 7, 63–73.
- Preparata, F.P., Frieze, A.M., Upfal, E., 1999a. On the power of universal bases in sequencing by hybridization. In: *Proceedings of Third Annual International Conference on Computers and Molecular Biology*, pp. 295–301.
- Preparata, F.P., Frieze, A.M., Upfal, E., 1999b. Optimal reconstruction of a sequence from its probes. *J. Comput. Biol.* 7 (3–4), 361–368.
- Sanger, F., Nickelen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74, 560–564.
- Shamir, R., Tsur, D., 2002. Large scale sequencing by hybridization. *J. Comput. Biol.* 9, 413–428.
- Southern, E.M., 1988. Analyzing polynucleotide sequences. *International Patent Application PCT/GB8900460*.
- Voss, S., Martello, S., Osman, I., Roucaïrol, C. (Eds.), 1998. *Meta-Heuristics—Advances and Trends in Local Search Paradigms for Optimization*. Kluwer Academic Publishers, Boston.
- Wallace, R.B., Johnson, M.J., Hirose, T., Miyake, T., Kawashima, E.H., Itakura, K., 1981. The use of synthetic oligonucleotides as hybridization probes. 2. Hybridization of oligonucleotides of mixed sequence to rabbit beta-globin DNA. *Nucleic Acids Res.* 9, 879–894.
- Zhang, J.-H., Wu, L.-Y., Zhang, X.-S., 2003. Reconstruction of DNA sequencing by hybridization. *Bioinformatics* 19, 14–21.