OPERATIONS RESEARCH
Vol. 54, No. 6, November–December 2006, pp. 1185–1192
ISSN 0030-364X | EISSN 1526-5463 | 06 | 5406 | 1185

informs

DOI 10.1287/opre.1060.0324

DNA Sequencing by Hybridization via Genetic Search

Jacek Blazewicz
Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań,
Poland, and the Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznań, Poland,
jblazewicz@cs.put.poznan.pl

Ceyda Oguz
Department of Industrial Engineering, Koç University, Istanbul, Turkey, coguz@ku.edu.tr

Aleksandra Swiercz, Jan Weglarz
Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań,
Poland, and the Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznań, Poland,
{aswiercz@cs.put.poznan.pl, jweglarz@cs.put.poznan.pl}

An innovative approach to DNA sequencing by hybridization utilizes isothermic oligonucleotide libraries. In this paper, we demonstrate the utility of a genetic algorithm for the combinatorial portion of this new approach by incorporating characteristics of DNA sequencing by hybridization in addition to isothermic oligonucleotide libraries. Specialized crossover and mutation operators were developed for this purpose. After initial experiments for parameter adjustment, the performance of the genetic algorithm approach was evaluated with respect to previous methods in the literature. The results indicate that the proposed new approach is superior to previous approaches. The proposed new crossover operator that inherits some features of the structured weighted combinations might also be of value for some other combinatorial problems, including the traveling salesman problem.

Subject classifications: analysis of algorithms: metaheuristics; health care: bioinformatics, DNA sequencing; programming: integer; algorithms: heuristic.
Area of review: Computing and Information Technologies.
History: Received May 2004; revisions received March 2005, September 2005; accepted October 2005.

1. Introduction

Advancing the scientific understanding of organisms through computation has an ever-increasing importance that necessitates more research into fundamental problems such as sequence analysis. Sequence analysis of DNA, in particular, has become an increasingly essential issue over the past few years, as we can see that nearly every research project in molecular biology involves sequence analysis and comparison (Schulze-Kremer 1995, Abbas and Holmes 2004). Furthermore, the ultimate goal of the Human Genome Project was to sequence accurately the entire genome, which from the computational point of view includes three main parts: *mapping*, *assembling*, and *sequencing* (Waterman 1995, Setubal and Meidanis 1997, Gusfield 1997, Pevzner 2000, Fogel and Corne 2003). All parts of the approach to find the genome require better analytical tools and more efficient algorithms because the volume of data is massive. The impact of the successful completion of the genome project is enormous and this alone proves the importance and value of developing more efficient and accurate algorithms for DNA sequencing. To this end, new technologies for DNA sequencing are emerging with the advance of technology, and they include sequencing by hybridization, among others (Human Genome Program 2003).

Considering the limitations and the shortcomings of the existing research that will be described in the next section, our attention in this paper is focused on developing an efficient, effective, and accurate heuristic algorithm for the combinatorial part of the DNA sequencing by hybridization (SBH) under the novel approach of isothermic libraries. This new method (Blazewicz et al. 1999a, 2004c) uses oligonucleotide libraries of equal melting temperatures (thus, differing by their lengths) and, as explained in the next section, could lead to overcoming one of the drawbacks of the classical SBH approach (with equal length oligonucleotide library), which sometimes cannot handle excessive rate of errors being a result of the biochemical phase of the SBH approach. It is, thus, of crucial importance to develop sequencing algorithms which will take into account the specificity of the data and will construct sequences as close as possible to the original ones. Based on promising results from the literature and in view of the success of metaheuristics in tackling difficult combinatorial problems, we considered developing a genetic algorithm (GA) taking the combinatorial nature of the problem into account. The obtained results, as demonstrated by an extensive computational experiment conducted on real DNA sequences, prove a clear superiority of the presented approach over the existing ones. Furthermore, tak-

ing into account the specificity of the data in this case (oligonucleotides of different length), this approach might also be of value when designing new algorithms for the assembling stage of genome reading. The proposed new crossover operator that inherits some features of the structured weighted combinations might also be of value for some other combinatorial optimization problems for which the solution can be represented as a permutation, such as the traveling salesman problem (TSP).

The rest of this paper is organized as follows. We overview the DNA sequencing by hybridization method in the next section. We then describe our proposed GA in §3. We next analyze the performance of our GA through its efficiency, effectiveness, and accuracy by extensive computational experiments in §4. We conclude the paper in §5.

2. DNA Sequencing by Hybridization—Formulation and Basic Properties

One of the significant breakthroughs in molecular biology was DNA sequencing, that is, the process of establishing the precise order of the bases along one strand of a DNA molecule. There are four different types of bases in a DNA molecule: *adenine*, *cytosine*, *guanine*, and *thymine*, which are abbreviated as A, C, G, and T, respectively. SBH, which is a third-generation gel-less technology for reading DNA, is one of the methods for DNA sequencing and is expected to be used more and more in the future (Human Genome Program 2001, Southern 1988, Drmanac et al. 1989, Blazewicz et al. 1999b, Blazewicz et al. 2000, Ben-Dor et al. 2001, Hubbell 2001, Phan and Skiena 2001, Blazewicz et al. 2002, Halperin et al. 2002, Shamir and Tsur 2002, Zhang et al. 2003, Blazewicz et al. 2004a). SBH starts with a biochemical experiment (*hybridization*). In the experiment, an unknown fragment of the single stranded DNA labeled either fluorescently or radioactively is hybridized to a DNA chip that holds the oligonucleotide library to be used. An oligonucleotide library is a collection of all possible oligonucleotides (*probes*), that is, short but known sequences of a few (2–10) nucleotides. The result of this process is a set of labeled oligonucleotides on a fluorescent or radioactive image of the DNA chip (*spectrum*). In other words, spectrum represents all overlapping oligonucleotides constituting the target DNA fragment (complementary to the labeled ones on the DNA chip). This spectrum, in turn, becomes the input data for the computational phase of SBH during which the sequence of the target DNA fragment is reconstructed. We stress here that these algorithms, after necessary adjustments and improvements, constitute the core of the *assembling stage* (being purely a computational approach) of every genome reconstruction (Venter et al. 2001; cf. also Blazewicz et al. 2004d, where the corresponding algorithm for the SARS co-virus genome assembling has been described). This motivates further their design and analysis.

One can observe that the success of SBH depends on the reliability of hybridization as well as on the efficiency and the accuracy of the algorithms used to reconstruct the sequence from the spectrum. In the classical SBH, an *isometric oligonucleotide library* of size 4^l is used to obtain the spectrum where each probe has equal length of l (Bains and Smith 1988, Fodor et al. 1991, Southern et al. 1992, Pevzner and Lipshutz 1994). If the hybridization experiment is an ideal one without any experimental errors, then one can reconstruct an original DNA sequence in polynomial time (Pevzner 1989) (see Example 1 in the online supplement at http://or.pubs.informs.org/pages/collect.html). In contrast, if the hybridization experiment results in some errors, then the problem handled in the computational phase becomes NP-hard (Blazewicz and Kasprzak 2003, Gallant et al. 1980). The errors generated during the hybridization can be of two types: positive errors occur when oligonucleotides that do not constitute the original sequence are included in the spectrum, and negative errors take place when oligonucleotides that are part of the original sequence are missing from the spectrum. A repetition—that is, any oligonucleotide appearing more than once in the original sequence—can be considered as a negative error because it will be included in the spectrum only once (see Example 2 in the online supplement at http://or.pubs.informs.org/Pages/collect.html).

The computational phase of the SBH approach with some types of errors appearing in the spectrum was addressed in few papers (Drmanac et al. 1989, Pevzner 1989, Bains 1991, Guénoche 1992, Lipshutz 1993, Blazewicz et al. 1999b, Phan and Skiena 2001, Halperin et al. 2002, Zhang et al. 2003). Very few of them dealt with unconstrained sets of errors; the most general approach is probably presented in Blazewicz et al. (1999b). In this approach, the computational phase of SBH has been reduced to a variant of the selective TSP with a nice mathematical programming formulation. Although this method is very general, an excessive number of errors in the spectrum makes this approach rather slow. Hence, a challenging problem is to design a new approach reducing the number of errors in the biochemical phase of SBH. Such an approach has been proposed in Blazewicz et al. (1999a).

The key idea of this new approach is to obtain a set of oligonucleotides that differ in base composition and length and are characterized by a predefined relation between the base composition and the length of oligonucleotides. In a specific case, if in a library the 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*. In what follows, the sum of increments of nucleotides forming an oligonucleotide will be called the oligonucleotide *temperature*.

The isothermic oligonucleotide library follows the experimentally established relationship between the base composition and the duplex stability. 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 character-

istic 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. The use of such libraries should substantially limit the number of these errors to be considered in the computational phase of the SBH approach. Moreover, these libraries will also avoid some repetitions (what follows different lengths of the oligonucleotides used in the new chip libraries). These repetitions are a serious drawback of the standard approach, thus, their elimination is a crucial step toward a practical application of the SBH approach. We give a formal reasoning for isothermic libraries in the online supplement and more details can be found in Blazewicz et al. (2004c).

Now we will formulate the isothermic sequencing problem in its most general form, i.e., with positive and negative errors. The formulation will be valid under the reasonable assumption that most of the data coming from the hybridization experiment are correct. Likewise in the classical SBH, as a result of the biochemical phase, one gets a set of oligonucleotides that hybridized with the unknown DNA sequence, i.e., spectrum ($\mathcal{S}$). Now the spectrum contains the data from the hybridization experiment with two isothermic oligonucleotide libraries differing by one increment of A(T) nucleotide. This problem in a search version can be viewed as the one of finding, for a given spectrum $\mathcal{S}$, 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 the maximization of the number of l-mers from the spectrum used to build a solution (a reconstructed sequence). It is assumed here that the only information provided by the hybridization experiment are the spectrum $\mathcal{S}$ and the length n of the DNA sequence that is looked for. Now we can define our problem as follows, where the necessary notation is given in Table 1.

Isothermic DNA sequencing with negative and positive errors—search version:

Instance: set $\mathcal{S}$ (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 $\beta + |\mathcal{S}| - 2\alpha$.

Its mathematical programming formulation is as follows (cf. Blazewicz et al. 2004b). Obviously, the solution of this mathematical programming formulation can be uniquely translated into a sequence of nucleotides. All variables appearing in the formulation are nonnegative integers.

Minimize

$$\beta + |\mathcal{S}| - 2\alpha \tag{1}$$

subject to

$$\sum_{i=1}^{|\mathcal{S}|} b_{ik} \leqslant 1, \quad k = 1, \ldots, |\mathcal{S}|, \tag{2}$$

Table 1. Notation.

α	number of oligonucleotides from the spectrum being a part of the constructed sequence.
β	number of oligonucleotides being members of the two used isothermic libraries (of temperatures equal to t and $t+2$, respectively), which can be distinguished in the sequence (each oligonucleotide adds to β the number of its occurrences in the sequence).
$\beta - \alpha$	number of negative errors.
$\lvert\mathcal{S}\rvert - \alpha$	number of positive errors.
s_i	element of the spectrum.
$s_i[j]$	jth nucleotide of s_i.
l_i	length of s_i.
n	length of an original sequence.
b_{ij}	Boolean variable; equal to 1 if element s_i is an immediate predecessor of element s_j in a reconstructed sequence, otherwise equal to 0.
c_{ij}	cost of joining element s_i (as the first one) with element s_j assuming a maximal overlap of the two elements (in the sense of the hybridization); equal to the difference between starting positions of the elements in a sequence obtained in the above way; if the difference is equal to zero and s_i is longer than s_j, the value of c_{ij} should be set to n.
y_{ijk}	Boolean variable; equal to one if in a sequence created from elements s_i (as first) and s_j joined with shift c_{ij}, it is possible to distinguish an oligonucleotide of temperature t starting from position k of the sequence; otherwise equal to zero.
y'_{ijk}	Boolean variable; equal to one if in a sequence created from elements s_i (as first) and s_j joined with shift c_{ij}, it is possible to distinguish an oligonucleotide of temperature $t+2$ starting from position k of the sequence; otherwise equal to zero.
y_{last}	number of oligonucleotides of temperatures t or $t+2$ possible to distinguish in the last element of the current reconstructed sequence.
$f(x)$	function returning the increment of nucleotide x, i.e., $f(\text{A}) = f(\text{T}) = 2$, $f(\text{C}) = f(\text{G}) = 4$.

$$\sum_{k=1}^{|\mathcal{S}|} b_{ik} \leqslant 1, \quad i = 1, \ldots, |\mathcal{S}|, \tag{3}$$

$$\sum_{k=1}^{|\mathcal{S}|} \left(\left| \sum_{i=1}^{|\mathcal{S}|} b_{ki} - \sum_{j=1}^{|\mathcal{S}|} b_{jk} \right| \right) = 2, \tag{4}$$

$$\sum_{s_k \in \mathcal{S}^*} \left(\sum_{s_i \in \mathcal{S}^*} b_{ik} \cdot \sum_{s_j \in \mathcal{S}^*} b_{kj} \right) < |\mathcal{S}^*| \quad \forall\, \mathcal{S}^* \subset \mathcal{S}, \mathcal{S}^* \neq \varnothing, \tag{5}$$

$$\text{last} = \frac{1}{2} \left[\sum_{k=1}^{|\mathcal{S}|} \left(\left| \sum_{i=1}^{|\mathcal{S}|} b_{ik} - \sum_{j=1}^{|\mathcal{S}|} b_{kj} \right| k \right) + \sum_{k=1}^{|\mathcal{S}|} \left(\left(\sum_{i=1}^{|\mathcal{S}|} b_{ik} - \sum_{j=1}^{|\mathcal{S}|} b_{kj} \right) k \right) \right], \tag{6}$$

$$\sum_{i=1}^{|\mathcal{S}|} \sum_{j=1}^{|S|} c_{ij} b_{ij} \leqslant n - l_{\text{last}}, \tag{7}$$

$$\alpha = \sum_{i=1}^{|\mathcal{S}|} \sum_{j=1}^{|\mathcal{S}|} b_{ij} + 1, \tag{8}$$

$$\beta = \sum_{i=1}^{|\mathcal{S}|} \sum_{j=1}^{|\mathcal{S}|} \sum_{k=1}^{c_{ij}} b_{ij}(y_{ijk} + y'_{ijk}) + y_{\text{last}}, \tag{9}$$

$$y_{ijk} \leqslant 1, \quad i = 1, \ldots, |\mathcal{S}|,\ j = 1, \ldots, |\mathcal{S}|,\ k = 1, \ldots, c_{ij}, \tag{10}$$

$$y'_{ijk} \leqslant 1, \quad i = 1, \ldots, |\mathcal{S}|,\ j = 1, \ldots, |\mathcal{S}|,\ k = 1, \ldots, c_{ij}, \tag{11}$$

$$y_{\text{last}} = y_{\text{last}\,1} + y_{\text{last}\,2} + 1, \tag{12}$$

$$y_{\text{last}\,i} \leqslant 1, \quad i = 1, 2, \tag{13}$$

$$\sum_{i=1}^{l_{\text{last}}-1} f(s_{\text{last}}[i]) = t \ \Leftrightarrow\ y_{\text{last}1} = 1, \tag{14}$$

$$\sum_{i=1}^{l_{\text{last}}-1} f(s_{\text{last}}[i+1]) = t \ \Leftrightarrow\ y_{\text{last}\,2} = 1, \tag{15}$$

$$\bigvee_{w=k}^{c_{ij}+l_j} \left(\sum_{z=k}^{w} f(s_{ij}[z]) = t \right) \ \Leftrightarrow\ y_{ijk} = 1, \quad i = 1, \ldots, |\mathcal{S}|,\ j = 1, \ldots, |\mathcal{S}|,\ k = 1, \ldots, c_{ij}, \tag{16}$$

$$\bigvee_{w=k}^{c_{ij}+l_j} \left(\sum_{z=k}^{w} f(s_{ij}[z]) = t+2 \right) \ \Leftrightarrow\ y'_{ijk} = 1, \quad i = 1, \ldots, |\mathcal{S}|,\ j = 1, \ldots, |\mathcal{S}|,\ k = 1, \ldots, c_{ij}, \tag{17}$$

$$z \leqslant l_i \ \Rightarrow\ s_{ij}[z] = s_i[z], \quad i = 1, \ldots, |\mathcal{S}|,\ j = 1, \ldots, |\mathcal{S}|,\ z = 1, \ldots, c_{ij} + l_j, \tag{18}$$

$$z > l_i \ \Rightarrow\ s_{ij}[z] = s_j[z - c_{ij}], \quad i = 1, \ldots, |\mathcal{S}|,\ j = 1, \ldots, |\mathcal{S}|,\ z = 1, \ldots, c_{ij} + l_j. \tag{19}$$

The minimization of criterion function (1) corresponds to the minimization of the number of errors connected with a reconstructed sequence. Inequalities (2) and (3) guarantee that each element of the spectrum has at most one immediate predecessor and at most one immediate successor in a reconstructed sequence. Equation (4) ensures that in a reconstructed sequence there will be only two elements not connected with other elements of the spectrum on both its ends. Inequalities (5) guarantee that there will be no cycle in a reconstructed sequence. Equation (6) assigns the index of the last element in the current reconstructed sequence. Inequality (7) ensures that a constructed sequence will not exceed length n. Constraints (8)–(19) define the values of parameters α and β.

The above problem is proved to be strongly NP-hard (Blazewicz et al. 2004b), thus, unlikely to admit polynomial- and pseudopolynomial-time optimization algorithms. There is also no hope of finding a fully polynomial approximation scheme. Hence, considering the success of metaheuristics in tackling the difficult combinatorial problems and recent applications of the hybrid GA and the tabu search algorithm to DNA sequencing with isometric and isothermic libraries, respectively, we developed a GA for DNA sequencing by using a hybridization approach with isothermic oligonucleotide libraries. The GA was especially suited for this purpose because it allowed us to incorporate problem-specific characteristics into its design and to handle different types of errors and large-sized problems.

3. Genetic Algorithm for the Isothermic SBH

In the last 15 years, there has been an increasing interest in metaheuristics for optimization problems. One of the reasons for this is that these methods provide one of the best ways to obtain a near optimal solution with a reasonable computational effort. Another reason is that they are designed for complex optimization problems where classical heuristic approaches and optimization methods cannot provide an efficient and effective solution. Metaheuristics are shown to work exceptionally well in practice (Rayward-Smith et al. 1996).

GA is one of the metaheuristics, and it involves more general and abstract techniques compared to other metaheuristics (Fraser 1957, Holland 1975). GA is a variation of evolutionary computation algorithms and specifically, it is a mechanism that simulates natural evolutionary processes. We refer the reader to Michalewicz (1996), Gen and Cheng (1997), and Reeves (2003) for a detailed description of the GAs.

Our proposed GA adopts the general structure of this metaheuristic with its standard components. However, we implemented each of these components in our GA by considering the characteristics of the DNA sequencing by using a hybridization approach with isothermic oligonucleotide libraries as explained in detail in the online supplement.

The input data for our GA are the spectrum $\mathcal{S}$, containing $|\mathcal{S}|$ oligonucleotides (possibly hybridizing with the sequence we look for) and the length n of the sequence, which are obtained from the biochemical phase of the isothermic SBH. In the following, we define o_f and o_l as the first and last oligonucleotides in an individual, respectively.

Step 1 (*initial population*). Create s individuals, where s is half of the sequence length. Each individual is a permutation of integers from 1 to $|\mathcal{S}|$.

Step 2 (*main loop*). Repeat Steps 3–5 until there are no improvements for iter $= 50$ generations.

Step 3 (*parent selection*). Define the fitness of an individual as the number of oligonucleotides from the spectrum used to form the best subsequence, which is no longer than n. Evaluate fitness of all individuals in the population and select $c * s$ parents according to the part-sum selection procedure, where $c = 0.9$. Steps 3.1 and 3.2, which describe the crossover operator, are repeated for each pair of parents, which are paired randomly.

Step 3.1 (*beginning of crossover*). Select the first oligonucleotide o_i randomly. Let $o_f = o_l = o_i$.

Step 3.2 (*crossover*). Find predecessor of o_f and successor of o_l in both parents. For all the oligonucleotides found, exclude the oligonucleotides that are already in the solution, and choose the one that fits better. If there is a tie, select one randomly. If there are no predecessors nor successors that are not used in the solution, find the oligonucleotide that fits the best, where ties are broken randomly. Assign the chosen oligonucleotide to the appropriate position in the solution. Update o_f and o_l. Repeat this step until all oligonucleotides are in the individual.

Step 4 (*mutation*). Choose an individual randomly (from both the offspring and the parents' populations). Find the oligonucleotide with the smallest total overlap degree and swap this oligonucleotide with the adjacent oligonucleotide that has the lowest overlap degree with it. Repeat this step with mutation frequency $m * s * |\mathcal{S}|$, where $m = 0.001$.

Step 5 (*creation of the next generation*). Evaluate fitness for each individual from the offspring and the parents' population. Select all the individuals from the offspring population and the best ones from the parents' population as the next generation.

The pattern of the crossover operator used in our approach inherits some features of *structured weighted combinations* (Glover 1994). One might find the similarities on how the customary operations (for example, a simple crossover) are replaced with structured transformations of (sub)sequences that preserve specified discrete relationships and associated feasibility conditions. Following three properties of the (sub)sequences from which the weighted combinations are created, we define:

PROPERTY 1. Each (sub)sequence represents a set of precedence relationships between neighboring oligonucleotides in the individual for a particular decision.

PROPERTY 2. In a solution, all oligonucleotides are in the individual exactly once. Although the created sequence is longer than n, the best subsequence of length not greater than n is selected, which is what makes the solution feasible.

PROPERTY 3. A new individual is created according to the combination of different precedence relationships between neighboring oligonucleotides from two parents. There are defined rules to determine which precedence relationships will be chosen so that Properties 1 and 2 continue to hold.

This approach was chosen to prevent inadvertent destruction of good traits in the parents.

4. Computational Results

In this section, we present the evaluation of the performance of the GA; we describe the data and the tuning of the parameters in the online supplement. The solution obtained from the GA can be evaluated in two ways. First, during computations, while no information about the order of oligonucleotides is provided, the number of oligonucleotides composing the solution determines the quality of the solution. This statement originates from the fact that most of the oligonucleotides in the spectrum are correct. Otherwise, it would be impossible to find the original sequence. Second, once computations are over, the sequence generated by the algorithm is compared to the original one, which results in the similarity value. The similarity is quantified according to the Needelman-Wunsch algorithm (Needelman and Wunsch 1970; cf. also Waterman 1995 and Setubal and Meidanis 1997).

Based on the above criteria, we evaluated the performance of the proposed GA compared to those of the earlier algorithms used for DNA SBH problem. For this purpose, we considered the best algorithms proposed for the DNA SBH problem with both positive and negative errors where an isometric oligonucleotide library is used, namely, the tabu search (TS) algorithm by Blazewicz et al. 2004a, which uses a scatter search for diversification, and the algorithm by Zhang et al. (2003). We have also considered the TS algorithm proposed by Blazewicz et al. (2004c) for the same problem with an isothermic oligonucleotide library. In the following, each entry of Tables 2–4 summarizes the results of the tests for 40 different sequences obtained from the GenBank in the way described in the online supplement.

In Table 2, the results of four algorithms are compared. There are two measures for the evaluation of the obtained

Table 2. Similarity of the obtained sequence to the original one and the number of optimal solutions for different algorithms.

		Zhang et al. (2003)		Blazewicz et al. (2004a)		Blazewicz et al. (2004c)		Genetic algorithm		
n	Errors	Similarity [%]	Optimal solutions	Similarity [%]	Optimal solutions	Similarity [%]	Optimal solutions	Similarity [%]	Optimal solutions	Deviation σ [%]
200	±5%	100	90/90	99.9	40/40	85.2	8/40	99.9	39/40	0.36
	±20%	—	—	97.9	36/40	78.7	3/40	99.2	37/40	3.47
400	±5%	100	80/80	95.3	32/40	75.9	2/40	99.2	38/40	4.68
	±20%	—	—	89.4	21/40	69.7	0/40	99.2	36/40	4.68
500	±5%	—	—	95.3	32/40	75.6	3/40	99.8	39/40	1.15
	±20%	—	—	83.9	17/40	70.2	0/40	99.6	35/40	1.85
600	±5%	—	—	95.2	32/40	76.5	2/40	98.0	36/40	7.71
	±20%	—	—	80.5	15/40	68.8	0/40	98.0	32/40	9.19

sequence: the similarity of the obtained sequence to the original one and the number of optimal solutions found by the algorithm. (Note that in our tests we knew the original sequences because they were DNA sequences taken from GenBank.) Each entry of the columns under the "optimal solutions" heading is composed of two numbers, a and b, and it means that among b instances, the algorithm found a original sequences, with 100% similarity. The cases for which the results were not given by Zhang et al. (2003) are denoted by a dash.

The first algorithm, described in Zhang et al. (2003), is mostly designed for positive errors in addition to some repetitions that can occur in the sequence. However, it works well with a small rate of negative errors—up to 5%. This algorithm was not tested for instances with a higher rate of negative errors nor for longer sequences.

The method described in Blazewicz et al. (2004a) seems to work well for hard instances with a high rate of both types of errors. However, even though the similarity is quite high (for sequences of 200-nucleotide length, similarity is above 99%), the number of optimal solutions found decreases sharply with an increase in the length of the sequence; for sequences with 600 nucleotides, only 15 out of 40 optimal solutions were found. One should note that tested instances contain longer sequences and with ±20% error rate compared to the algorithm by Zhang et al. (2003) and because more oligonucleotides are missing in the spectrum, the sequencing problem is more difficult.

The TS algorithm of Blazewicz et al. (2004c) for isothermic sequencing is presented in columns 7 and 8 of Table 2. The original algorithm assumed knowledge of the first oligonucleotide. To keep the same conditions for a fair testing of different algorithms, we changed the algorithm accordingly so that we do not use this additional information, which is the case for our GA. The TS algorithm solved the instances with the similarity in range [68.8%–85.2%], and it rarely found the original sequence.

The results of our GA are presented in the last three columns of Table 2. The algorithm works very well for all the instances, giving a high similarity rate ranging from 98% in the worst case to almost 100% for sequences with not more than 500 nucleotides. Among the instances with 400-nucleotide sequences and ±20% of positive and negative errors, the algorithm found the solution for four of the instances with low similarity to the original sequence, and hence the similarity appears to be equal to 99.2%. As can be seen from the optimal solution value, the algorithm finds the optimal solutions for the remaining instances. For sequences of length 600, 98% similarity and 36 (for ±5% error rate) and 32 (for ±20% error rate) out of 40 optimal solutions are very good results.

In the last column of Table 2, we present an additional measure for the performance of the GA: the deviation of the similarity values from the average similarity value. It is defined as the square root of the variation according to the function

$$\sigma = \sqrt{\frac{\sum_{i=1}^{b}(x_i - \overline{x})^2}{b-1}},$$

where x_i are the consecutive values of the similarity, $\overline{x}$ is the mean value, and b is the number of tested instances ($b = 40$). One might notice that the deviation is very small. Only for the case where the length of the sequence is equal to 600 nucleotides does it rise up to 9.2%. This rather high value follows the reason that there are more than 30 optimal solutions out of 40 and the average similarity is only 98%, which makes 2% of difference for each of the optimal solutions. Moreover, there were also few sequences with a rather small similarity value, which was around 60%.

Tables 3 and 4 show a comparison of the usage of oligonucleotides and CPU time for the algorithms working with isothermic sequencing. For these two methods, the criterion function was the number of oligonucleotides used for composing the solution. Thus, it is very important to obtain high value of the usage. In Table 3, each entry is the mean value for all the results. The value 100% means that the number of oligonucleotides is the same as the number of proper oligonucleotides in the spectrum. In Table 4, the CPU time of computations of both methods is compared under the same conditions.

It is very significant that the GA works for a similar period of time for the same length of the sequence but different error rate. Although it works longer than the TS algorithm for the case with ±5%, it is much faster while the error rate increases, and the results are better in both cases (cf. Table 2). It might be due to the different functions used to compare the intermediate solutions in the algorithms. In the TS algorithm, the function for evaluating intermediate solutions was a condensation, i.e., the number of oligonucleotides divided by the length of the sequence. In the GA, we tested different functions (as explained in the online supplement), including the condensation, to be used within the crossover operator. Among the functions that were tested, the best one appeared to be the combination of overlaps between neighboring oligonucleotides and the temperature of the sequence. We see that it indirectly takes the length of oligonucleotides into consideration. This function provides very good results—almost all the obtained sequences are optimal from the criterion function point of

Table 3. Usage of oligonucleotides for isothermic sequencing.

	Blazewicz et al. (2004c)		Genetic algorithm	
n	±5%	±20%	±5%	±20%
200	96.8	96.1	100	100
400	96.6	95.3	100	99.9
500	96.0	95.6	100	99.9
600	96.2	95.1	100	99.9

Table 4. CPU time of computations [s].

	Blazewicz et al. (2004c)		Genetic algorithm	
n	±5%	±20%	±5%	±20%
200	3.5	9.7	6.5	8.5
400	18.0	74.7	23.9	30.8
500	30.0	125.5	46.5	53.6
600	45.9	199.9	80.9	91.6

view because the number of oligonucleotides is almost the same as the number of proper oligonucleotides in the spectrum. Moreover, they are also optimal from the biochemical point of view because similarity to original sequences is almost 100% (cf. Table 2). For the TS algorithm, the quality of the solution, although very high (95%–97%), does not lead to high similarity of the obtained sequence when compared to the original one.

Finally, we observe from the results that in the cases of 500-nucleotide and 600-nucleotide sequences with error rate ±20%, the GA sometimes could not find the original sequence. Even though usage of oligonucleotides is 100%, similarity is lower. The reason for this could be the existence of more than one solution derived from the same number of oligonucleotides. Without any additional (biochemical) information about the original sequence, it cannot be indicated which one is better because they are equivalent with respect to the function evaluated by the GA.

5. Conclusions

In this paper, the problem of DNA sequencing by hybridization with isothermic libraries is solved by the GA. This combination proved to be very efficient, outperforming other approaches in the field. Extensive tests made on real DNA sequences, which were taken from GenBank, have demonstrated the high quality of the solutions generated, measured by the similarity to the original sequences and the number of the optimal sequences found. As demonstrated by examples, this new approach can also cope with some repetitions, whereas the classical isometric approach cannot. However, there is still room for improvement, especially in cases of a high repetition rate in original sequences. Here, it seems to be promising to use the information from the problem structure in an intelligent way and to reduce the randomness in the GA.

One strength of GAs can be attributed to the crossover operator used that recombines good traits of two parents from a population of solutions. But an accompanying weakness of GAs is the inadvertent destruction of these good traits during the crossover. To alleviate this weakness, one can consider incorporating the strengths of scatter search and path relinking (Glover and Laguna 1997, Glover et al. 2000), which are also population-based metaheuristics, into the GA. In this study, we particularly considered a structured recombination of good traits of the individuals, such as weighted structured combinations used in scatter search and path relinking (Glover 1994). Such an approach has brought additional power to our GA by combining solutions more intelligently than the classical crossover operators (Glover 2004). This approach can also be of value for some other combinatorial problems for which the solution can be represented as a permutation, including the TSP.

Acknowledgments

The authors are grateful to Marta Kasprzak for comments on their work and for additional tests she provided. They are also grateful to the anonymous referees for their thorough comments, which greatly improved the presentation of the ideas and the results contained in this paper. The work described in this paper was undertaken when Ceyda Oguz was with the Hong Kong Polytechnic University and was partially supported by a grant from the Hong Kong Polytechnic University (project no. A-PE91) and by a KBN grant (No. 3T11F 002 27) from the Ministry of Science of Poland.

References

Abbas, A. E., S. P. Holmes. 2004. Bioinformatics and management science: Some common tools and techniques. *Oper. Res.* **52** 165–190.

Bains, W. 1991. Hybridization methods for DNA sequencing. *Genomics* **11** 294–301.

Bains, W., G. C. Smith. 1988. A novel method for nucleic acid sequence determination. *J. Theoretical Biology* **135** 303–307.

Ben-Dor, A., I. Pe'er, R. Shamir, R. Sharan. 2001. On the complexity of positional sequencing by hybridization. *J. Comput. Biol.* **6** 361–371.

Blazewicz, J., M. Kasprzak. 2003. Complexity of DNA sequencing by hybridization. *Theoretical Comput. Sci.* **290** 1459–1473.

Blazewicz, J., F. Glover, M. Kasprzak. 2004a. DNA sequencing—Tabu and scatter search combined. *INFORMS J. Comput.* **16** 232–240.

Blazewicz, J., M. Kasprzak, W. Kuroczycki. 2002. Hybrid genetic algorithm for DNA sequencing with errors. *J. Heuristics* **8** 495–502.

Blazewicz, J., P. Formanowicz, M. Kasprzak, W. T. Markiewicz. 1999a. Method of sequencing of nucleic acids. Polish patent application P335786.

Blazewicz, J., P. Formanowicz, M. Kasprzak, W. T. Markiewicz. 2004b. Sequencing by hybridization with isothermic oligonucleotide libraries. *Discrete Appl. Math.* **145** 40–51.

Blazewicz, J., P. Formanowicz, M. Kasprzak, W. T. Markiewicz, A. Swiercz. 2004c. Tabu search algorithm for DNA sequencing by hybridization with isothermic libraries. *Comput. Biol. Chemistry* **28** 11–19.

Blazewicz, J., P. Formanowicz, M. Kasprzak, W. T. Markiewicz, J. Weglarz. 1999b. DNA sequencing with positive and negative errors. *J. Comput. Biol.* **6** 113–123.

Blazewicz, J., P. Formanowicz, M. Kasprzak, W. T. Markiewicz, J. Weglarz. 2000. Tabu search for DNA sequencing with false negative and false positives. *Eur. J. Oper. Res.* **125** 257–265.

Blazewicz, J., M. Figlerowicz, P. Formanowicz, M. Kasprzak, B. Nowierski, R. Styszynski, L. Szajkowski, P. Widera, M. Wiktorczyk. 2004d. Assembling SARS-CoV genome—A new method based on graph theoretic approach. *Acta Biochimica Polonica* **51**(4) 983–993.

Drmanac, R., I. Labat, I. Brukner, R. Crkvenjakov. 1989. Sequencing of megabase plus DNA by hybridization: Theory and method. *Genomics* **4** 114–128.

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

Fogel, G. B., D. W. Corne, eds. 2003. *Evolutionary Computations in Bioinformatics*. Morgan Kaufman, San Francisco, CA.

Fraser, A. S. 1957. Simulation of genetic systems by automatic digital computers. I. Introduction. *Australian J. Biol. Sci.* **10** 484–491.

Gallant, J., D. Maier, J. A. Storer. 1980. On finding minimal length superstrings. *J. Comput. System Sci.* **20** 50–58.

Gen, M., R. Cheng. 1997. *Genetic Algorithms and Engineering Design*. John Wiley and Sons, New York.

Glover, F. 1994. Tabu search for nonlinear and parametric optimization (with links to genetic algorithms). *Discrete Appl. Math.* **49** 231–255.

Glover, F. 2004. Private communication.

Glover, F., M. Laguna. 1997. *Tabu Search.* Kluwer Academic Publishers, Boston, MA.

Glover, F., M. Laguna, R. Martí. 2000. Fundamentals of scatter search and path relinking. *Control and Cybernetics* **29** 653–684.

Guénoche, A. 1992. Can we recover a sequence, just knowing all its subsequences of given length? *CABIOS— Comput. Appl. Bioscience* **8** 569–574.

Gusfield, D. 1997. *Algorithms on String, Trees and Sequences. Computer Science and Computational Biology*. Cambridge University Press, New York.

Halperin, E., S. Halperin, T. Hartman, R. Shamir. 2002. Handling long targets and errors in sequencing by hybridization. *Proc. 6th Ann. Internat. Conf. Res. Comput. Molecular Biology* (RECOMB), Washington D.C., 176–185.

Holland, H. 1975. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI.

Hubbell, E. 2001. Multiplex sequencing by hybridization. *J. Comput. Biol.* **8** 141–149.

Human Genome Program. 2003. *Genomics and Its Impact on Medicine and Society: A 2003 Primer*. http://www.ormt.gov/ sci/ techresources/ Human_Genome/publicat/primer2004/index.shtml. U.S. Department of Energy.

Lipshutz, R. J. 1993. Likelihood DNA sequencing by hybridization. *J. Biomolecular Structural Dynamics* **11** 637–653.

Michalewicz, Z. 1996. *Genetic Algorithms + Data Structures = Evolution Programs*, 3rd rev. and extended ed. Springer-Verlag, Berlin, Germany.

Needelman S. B., C. D. Wunsch. 1970. A general method applicable to the search for similarities of the amino acid sequence of two proteins. *J. Molecular Biol.* **48** 443–453.

Pevzner, P. A. 1989. *l*-tuple DNA sequencing: Computer analysis. *J. Biomolecular Structure and Dynamics* **7** 63–73.

Pevzner, P. A. 2000. *Computational Molecular Biology: An Algorithmic Approach*. MIT Press, Cambridge, MA.

Pevzner, P. A., R. J. Lipshutz. 1994. Towards DNA sequencing chips. I. Prívara, I. Rovan, B. Ruicka, eds. *Proc. 19th Internat. Sympos. Math. Foundations of Comput. Sci.* Springer-Verlag, Berlin, Germany, 143–158.

Phan, V. T., S. Skiena. 2001. Dealing with errors in interactive sequencing by hybridization. *Bioinformatics* **17** 862–870.

Rayward-Smith, V. J., I. H. Osman, C. R. Reeves, G. D. Smith, eds. 1996. *Modern Heuristic Search Methods*. John Wiley, Chichester, UK.

Reeves, C. R. 2003. Genetic algorithms. F. Glover, G. A. Kochenberger, eds. *Handbook of Metaheuristics*, Vol. 3. Kluwer Academic Publishers, Boston, MA, 55–82.

Schulze-Kremer, S. 1995. *Molecular Bioinformatics: Algorithms and Applications*. de Gruyter, Berlin, Germany.

Setubal, J., J. Meidanis. 1997. *Introduction to Computational Molecular Biology*. PWS Publishing Company, Boston, MA.

Shamir, R., D. Tsur. 2002. Large scale sequencing by hybridization. *J. Comput. Biol.* **9** 413–428.

Southern, E. M. 1988. United Kingdom patent application GB8810400.

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

Venter, J. C. et al. 2001. The sequence of the human genome. *Science* **291** 1304–1351.

Waterman, M. S. 1995. *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman and Hall, London, UK.

Zhang, J.-H., L.-Y. Wu, X.-S. Zhang. 2003. Reconstruction of DNA sequencing by hybridization. *Bioinformatics* **19** 14–21.