Journal of Heuristics, 8: 495–502, 2002

Hybrid Genetic Algorithm for DNA Sequencing with Errors*

JACEK BŁAŻEWICZ[†] AND MARTA KASPRZAK[‡]
Institute of Computing Science, Poznań University of Technology, Poland; Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland
email: blazewic@put.poznan.pl

WOJCIECH KUROCZYCKI
Institute of Computing Science, Poznań University of Technology, Poland

Abstract

In the paper, a new hybrid genetic algorithm solving the DNA sequencing problem with negative and positive errors is presented. The algorithm has as its input a set of oligonucleotides coming from a hybridization experiment. The aim is to reconstruct an original DNA sequence of a known length on the basis of this set. No additional information about the oligonucleotides nor about the errors is assumed. Despite that, the algorithm returns for computationally hard instances surprisingly good results, of a very high similarity to original sequences.

Key Words: genetic algorithms, DNA sequencing by hybridization, negative and positive errors

1. Biochemical preliminaries and problem formulation

The *DNA sequencing* is one of the most important problems in the computational molecular biology. Its aim is to determine a sequence of nucleotides of an examined DNA fragment. A DNA fragment is usually written as a sequence of letters A, C, G, and T, representing the four nucleotides composing the fragment, i.e. adenine, cytosine, guanine, and thymine, respectively. A short sequence of nucleotides is called an *oligonucleotide*. The sequencing process uses as input data a set of oligonucleotides of equal length, being subsequences of one strand of the examined DNA fragment, and coming from a hybridization experiment. Next, an original sequence of a known length should be reconstructed on the basis of the oligonucleotides, overlapping one another.

In the *hybridization experiment* (Bains and Smith, 1988; Lysov et al., 1988; Southern, 1988; Drmanac et al., 1989; Markiewicz et al., 1994), a complete oligonucleotide library is compared with many copies of one strand of the examined DNA fragment. The library consists of all (4^l) short one-strand DNA fragments of length l. In order to identify fragments from the library, they are constructed in a special way on a *DNA chip* (Southern, 1988; Fodor

*The research has been partially supported by KBN grant 7T11F02621.
[†] Author to whom all correspondence should be addressed.
[‡] Fellowship holder of the Foundation for Polish Science.

et al., 1991; Caviani Pease et al., 1994), each element of the library having unique coordinates of the chip. During the hybridization reaction, copies of the longer DNA fragment join to oligonucleotides from the library in their complementary places. After the reaction, reading a fluorescent image of the chip one obtains the set of oligonucleotides being subfragments of the examined DNA fragment. This set is named *spectrum*.

If the hybridization experiment was carried on without any errors, then the spectrum would be *ideal*, i.e. it would contain only all subsequences of length l of the original sequence of the known length n. In this case the spectrum would consist of $n - l + 1$ elements and to reconstruct the original sequence one should find the order of spectrum elements such that neighboring elements always overlap on $l - 1$ nucleotides (see Example 1). There are several exact methods solving the DNA sequencing problem with the ideal spectrum, described for example in Bains and Smith (1988), Lysov et al. (1988), or in Drmanac et al. (1989), but only the one proposed in Pevzner (1989), and based on the transformation to the Eulerian path problem, works in polynomial time.

Example 1. Let the original sequence to be found is ACTCTGG, $n = 7$. In the hybridization experiment we can use, for example, the complete library of oligonucleotides of length $l = 3$. It is composed of the following $4^3 = 64$ oligonucleotides: {AAA, AAC, AAG, AAT, ACA, ..., TTG, TTT}. As a result of the hybridization experiment performed without experimental errors, one gets the ideal spectrum for this sequence, containing all 3-letters substrings of the original sequence: {ACT, CTC, CTG, TCT, TGG}. The reconstruction of the sequence consists in finding such an order of the spectrum elements, that each pair of neighboring elements overlap on $l - 1 = 2$ letters. The only possible solution for the example is presented in figure 1.

However, the hybridization experiment usually produces errors in the spectrum. There are two types of *errors: negative* ones, i.e. missing oligonucleotides in the spectrum, and *positive* ones, being erroneous oligonucleotides. The presence of negative errors forces the overlapping between some neighboring in a sequence oligonucleotides on less than $l - 1$ letters. The presence of positive errors in the spectrum forces a rejection of some oligonucleotides during the reconstruction process. The existence of errors in the DNA sequencing results in the strongly NP-hard combinatorial problems (Błażewicz and Kasprzak, 2002). There exist methods assuming errors in the spectrum, exact and heuristic ones, but almost all of them consider a reduced model of errors: Pevzner (1989), Drmanac et al. (1991), Lipshutz (1993), Hagstrom et al. (1994), and Błażewicz et al. (1997). The only exact method for the DNA sequencing problem allowing for any type of errors and no additional information about the spectrum, has been presented in Błażewicz et al. (1999a). It generates solutions

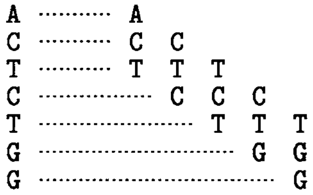

Figure 1. The reconstruction of the original sequence from the ideal spectrum.

composed of a maximal number of spectrum elements (a version of the Selective Traveling Salesman Problem), what leads to the reconstruction of original sequences (see Example 2). The same criterion function has been used in the tabu search methods for the problem with the most general model of errors (Błażewicz et al., 1999b, 2000).

Example 2. To make the problem of the *DNA* sequencing on the basis of the spectrum from Example 1 computationally hard, we introduce to the spectrum some errors. Let the negative error be CTC, and the positive errors be CAA and TTG. Then, the spectrum would have the following components: $\{$ACT, CAA, CTG, TCT, TGG, TTG$\}$. The usage of the criterion function from Błażewicz et al. (1999a), i.e. the maximization of the number of spectrum elements composing the solution of length not greater than $n = 7$, would produce here the two following orders of the oligonucleotides: (ACT, TCT, CTG, TGG) and (CAA, ACT, CTG, TGG), resulting in the two optimal solutions: ACTCTGG and CAACTGG, respectively. One of them is the original sequence. However, the data coming from real hybridization experiments, usually allow for a reconstruction of only one optimal solution.

In the paper, a new hybrid genetic algorithm solving the DNA sequencing problem with negative and positive errors, is presented. This algorithm supplements a standard genetic approach by using a heuristic greedy improvement. The original proposal to create combined solutions and to enhance them heuristically was in the scatter search method introduced in Glover (1977) (cf. Glover and Laguna, 1997). The hybrid genetic algorithm returns for computationally hard instances surprisingly good results, of a very high similarity to original sequences. The organization of the paper is as follows. In Section 2 the new algorithm is described. A comparison of the new method with the tabu search method proposed in Błażewicz et al. (1999b) is presented in Section 3. Section 4 concludes the paper.

2. The algorithm

The algorithm proposed in this paper uses the same *criterion function* as the previous methods solving the DNA sequencing problem with negative and positive errors. The goal is to maximize the number of elements from a spectrum composing a solution being a sequence of nucleotides not longer than n (it can be shorter in case of negative errors at the end). The criterion function is justified by the fact that most of the information from the hybridization experiment is correct. In the other case it would be impossible to reconstruct an original sequence without additional information, hard to obtain. The algorithm also accepts the *general model of errors*, i.e. it assumes the existence of negative and positive errors in the spectrum. Thus, as the input of the algorithm we have only the spectrum (an arbitrary set of words of equal length l) and a value of n. The main scheme of the algorithm is based on the idea of genetic algorithms (Holland, 1975; Goldberg, 1989).

The *genetic representation* of an individual (i.e. a *chromosome*) is a permutation of indices of oligonucleotides from the spectrum. The adjacency-based coding has been used: value i at position j in the chromosome means that the oligonucleotide i follows the oligonucleotide j. The function evaluating a fitness of an individual (the *fitness function*) takes the best substring of oligonucleotides in the chromosome, i.e. the one composed of the most elements, provided it produces a sequence of the length not greater than n nucleotides. The neighboring oligonucleotides are assumed to be maximally overlapped, what gives the guarantee of including in the evaluated substring as many elements as possible. The normalized fitness value, used in the algorithm, equals the number of oligonucleotides in this substring divided by $n - l + 1$ (being the maximum number of spectrum elements in any valid sequence).

The *initial population* is randomly generated according to the uniform distribution, and its cardinality s is a parameter of the method. Each of the individuals has to be the permutation of indices (as mentioned above) and it has to include no subcycle involving the lower number of indices than the spectrum cardinality. Next, to each individual the normalized fitness value is assigned. The individual of the greatest value of the criterion function is stored. Then, the fitness values of all individuals in the population are *linearly scaled*, and the best ones are selected according to the *stochastic remainder method without replacement* (Goldberg, 1989). The next population is constructed from the best individuals, randomly paired, using the *greedy crossover*, an approach similar to the one from Grefenstette et al. (1985) (see also Glover, 1977) in the context of a scatter search approach). The greedy crossover is defined as follows. The first oligonucleotide in a chromosome is set randomly. Next, with the probability 20% we choose for a given oligonucleotide in the chromosome the best successor among the remaining oligonucleotides (the ones not yet used to build the chromosome). As the best successor we understand the oligonucleotide which overlaps the previous one on the highest number of nucleotides. With the probability 80% the following move is chosen: if it does not produce a subcycle in the chromosome, we take as the successor for a given oligonucleotide the one with a better overlap in the parents of the chromosome, otherwise we take a random oligonucleotide among the remaining ones. In all the cases, if there is more than one best choice, the first found is chosen. The moves are done until all chromosomes of the population are constructed.

Every new population is submitted to the above series of operations, and each time the best individual found so far is remembered. The steps are repeated until a given number r of iterations without improvement of the criterion function value is reached. The solution returned by the algorithm is a part of the best individual found during the computations.

3. Computational results

In the computational experiment, the proposed algorithm has been compared with the tabu search method described in Błażewicz et al. (1999b) (its previous version being published in Błażewicz et al. (2000)). The tabu algorithm uses a greedy procedure generating initial solutions (Błażewicz et al., 1999a). Parameters of the tabu algorithm have been set to values resulting in similar computation times as used by the new algorithm. However, the ratio of condensing to extending moves (as defined for this tabu search algorithm), and a length of the tabu list, have been constant during all the tests. The parameters of the genetic algorithm have been set in a preliminary test to the following values: $r = 20$ and $s = 50$. These values led to an approximate optimization of two important criteria: the quality of solutions and the computation times.

The computational experiment has been performed on a PC station with Pentium II 300 MHz processor, 256 MB RAM and Linux operating system. All spectra used in the experiment have been derived from the DNA sequences coding human proteins (taken from GenBank, National Institute of Health, USA). They contain 20% of random negative errors and 20% of random positive errors (40% in total). Because cardinalities of the spectra vary from 100 to 500 oligonucleotides, they contain from 40 to 200 errors (in the latter case 100 randomly chosen oligonucleotides are missing and in addition 100 oligonucleotides in a spectrum are erroneous). The spectra have been sorted alphabetically, thus no information about an original order of oligonucleotides in sequences has been kept. The size of oligonucleotides in all the cases is equal to 10. The lengths of original sequences ($109 \leq n \leq 509$) and of oligonucleotides ($l = 10$) have been chosen on the basis of real hybridization experiments (cf. Caviani Pease et al., 1994). However, both compared algorithms accept any values of n and l, provided $l \leq n$.

The sequences produced by both methods have been compared with original sequences using a classical pairwise alignment algorithm (Waterman, 1995). The algorithm has been called with the following parameters: a match (the same nucleotides at a given position in strings) brings a profit of 1 point, a mismatch (different nucleotides) brings a penalty of 1 point (i.e. -1) and a gap (an insertion, a nucleotide against a space) also brings a penalty of 1 point. Therefore, the highest score (similarity) would be equal to a number of nucleotides in the sequences (in the case we consider the same two sequences) and the lowest score would be equal to a number of nucleotides in the longer sequence multiplied by -1 (in the case we have two completely different sequences).

In Table 1, computational results of the new algorithm are presented. All entries with average values have been calculated for 40 instances, derived from 40 different sequences. The quality means a number of spectrum elements composing a solution. For the given instances, a value of the criterion function reached by the algorithm cannot exceed the optimal quality, being the number of proper oligonucleotides in a spectrum. Below qualities, there are shown numbers of optimal solutions, among 40, returned by the algorithm. Similarity scores, summed up as described above, are shown as numbers of points (with maximal values from 109 to 509, respectively) and in percentages (with the maximum 100% in the case two sequences are equal).

The tests have proved the very good performance of the proposed algorithm. During short computation time it generates near-optimal solutions, and their similarities to original sequences are very high. For instances of cardinality 100, the algorithm returned only original

Table 1. Results of the genetic algorithm.

Spectrum size	100	200	300	400	500
Average quality	80.0	159.4	237.6	315.9	393.2
Optimal quality	80	160	240	320	400
Optimum no.	40	31	20	9	5
Average similarity score (pt)	108.4	199.3	274.1	301.7	326.0
Average similarity score (%)	99.7	97.7	94.3	86.9	82.0
Average computation time (sec)	13.5	63.4	154.9	263.4	437.9

Table 2. Results of the tabu search method.

Spectrum size	100	200	300	400	500
Average quality	80.0	158.6	235.5	313.8	391.1
Optimal quality	80	160	240	320	400
Optimum no.	40	24	11	6	2
Average similarity score (pt)	108.4	184.1	196.6	229.5	235.1
Average similarity score (%)	99.7	94.0	81.8	78.1	73.1
Average computation time (sec)	14.1	60.8	177.7	258.3	471.5

sequences. Similarity less than 100% is caused in that case only by missing information about last nucleotides in sequences (negative errors). Even for large spectra with many errors of both types, the algorithm composes very good (even optimal) sequences. The obtained solutions have the qualities from 98.3% to 100% of optimal values on the average. It should be noticed, that sometimes an instance has more than one optimal solution. In that case an optimal solution returned by the algorithm can differ from an original one. Thus, similarities presented in the table are in fact lower bounds on the quality measure of the algorithm.

For a comparison, Table 2 contains results produced by the tabu search method from Błażewicz et al. (1999b). The algorithm also returns for sets of cardinalities 100 as good solutions as possible. However, with a growing instance size the results become worse. The average qualities are also very close to the optimal ones (from 97.8% to 100%), but similarities of solutions to original sequences are lower than in the case of the algorithm presented here. It also should be pointed, that the proposed genetic algorithm returns many more optimal solutions than the tabu search method. It can happen, that a biochemist who would like to get the sequence reconstructed on the basis of his experiment, is interested in obtaining the exact solutions only. Of course, because the considered DNA sequencing problem with errors is strongly NP-hard, this is not always possible. Thus, the method working in polynomial time, often returning optimal solutions, is very valuable from the theoretical and practical points of view.

Other series of tests have been done to estimate how both methods work within longer computation time. Only the hardest instances have been chosen in order to observe a potential improvement. The time has been set to about 40 minutes. It required the change of the values of the genetic algorithm parameters to: $r = 40$ and $s = 200$. The tabu search method has been called with more local search steps and more restarts of the search procedure. The results of that experiment are presented in Table 3.

Table 3. Results of both methods, for spectra of cardinality 500, with computation time set to 40 min.

	GA method	TS method
Average quality	396.0	394.1
Optimum no.	9	4
Average similarity score (pt)	393.1	286.0
Average similarity score (%)	88.6	78.1

Again, the average qualities of solutions produced by both algorithms are very high and similar, and the similarity score for the new algorithm is much better than for the previous one. We conclude that the proposed algorithm has clear advantages. These results suggest the value of an evolutionary approach in this setting. An interesting possibility is the evolutionary scatter search method, which can be joined with tabu search, in order to combine the advantages of both types of procedures, or enhance the tabu search method by implementing the candidate list strategy and a stronger intensification strategy.

4. Conclusion

The results of the computational experiment have appeared to be very good, however, they can be further improved. The presented method has not used any additional information about spectra, which could be derived from biochemical experiments. For example, one can assume that first (or last) oligonucleotide of an original sequence is known. This assumption is based on the knowledge about primers, used by biochemists to amplify an examined molecule in PCR reaction before sequencing. Then, the obtained sequences would be much more similar to original ones. Another information may come from databases - a probabilistic analysis of existing characteristic subsequences in particular genes could exclude several low-probability orderings of oligonucleotides. However, the additional information is not always accessible, so we proposed a more general algorithm of a wider applicability. It should be noticed, that the ratio of errors to proper oligonucleotides in the tests is rather big. In real experiments one can expect lower number of errors and then the results produced by the algorithm would have been better.

References

Bains, W. and G.C. Smith. (1988). “A Novel Method for Nucleic Acid Sequence Determination.” *Journal of Theoretical Biology* 135, 303–307.

Błażewicz, J., J. Kaczmarek, M. Kasprzak, W.T. Markiewicz, and J. Węglarz. (1997). “Sequential and Parallel Algorithms for DNA Sequencing.” *Computer Applications in the Biosciences* 13, 151–158.

Błażewicz, J., P. Formanowicz, M. Kasprzak, W.T. Markiewicz, and J. Węglarz. (1999a). “DNA Sequencing with Positive and Negative Errors.” *Journal of Computational Biology* 6, 113–123.

Błażewicz, J., P. Formanowicz, F. Glover, M. Kasprzak, and J. Węglarz. (1999b). “An Improved Tabu Search Algorithm for DNA Sequencing with Errors.” In *Proceedings of the III Metaheuristics International Conference MIC’99*, pp. 69–75.

Błażewicz, J., P. Formanowicz, M. Kasprzak, W.T. Markiewicz, and J. Węglarz. (2000). “Tabu Search for DNA Sequencing with False Negatives and False Positives.” *European Journal of Operational Research* 125, 257–265.

Błażewicz, J. and M. Kasprzak. (2002). “Complexity of DNA Sequencing by Hybridization.” Theoretical Computer Science, to appear.

Caviani Pease, A., D. Solas, E.J. Sullivan, M.T. Cronin, C.P. Holmes, and S.P.A. Fodor. (1994). “Light-Generated Oligonucleotide Arrays for Rapid DNA Sequence Analysis.” In *Proceedings of the National Academy of Sciences of the USA* 91, pp. 5022–5026.

Drmanac, R., I. Labat, I. Brukner, and R. Crkvenjakov. (1989). “Sequencing of Megabase Plus DNA by Hybridization: Theory of the Method.” *Genomics* 4, 114–128.

Drmanac, R., I. Labat, and R. Crkvenjakov. (1991). “An Algorithm for the DNA Sequence Generation from k-tuple Word Contents of the Minimal Number of Random Fragments.” *Journal of Biomolecular Structure and Dynamics* 8, 1085–1102.

Fodor, S.P.A., J.L. Read, M.C. Pirrung, L. Stryer, A.T. Lu, and D. Solas. (1991). "Light-Directed, Spatially Addressable Parallel Chemical Synthesis." *Science* 251, 767–773.

Glover, F. (1977). "Heuristics for Integer Programming Using Surrogate Constraints." *Decision Sciences* 8, 156–166.

Glover, F. and M. Laguna. (1977). *Tabu Search*. Norwell: Kluwer Academic Publishers.

Goldberg, D.E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading: Addison-Wesley.

Grefenstette, J.J., R. Gopal, B.J. Rosmaita, and D. Van Gucht. (1985). "Genetic Algorithms for the Traveling Salesman Problem." In *Proceedings of International Conference on Genetic Algorithms and Their Applications*, pp. 160–168.

Hagstrom, J.N., R. Hagstrom, R. Overbeek, M. Price, and L. Schrage. (1994). "Maximum Likelihood Genetic Sequence Reconstruction from Oligo Content." *Networks* 24, 297–302.

Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor: University of Michigan Press.

Lipshutz, R.J. (1993). "Likelihood DNA Sequencing by Hybridization." *Journal of Biomolecular Structure and Dynamics* 11, 637–653.

Lysov, Yu. P., V.L. Florentiev, A.A. Khorlin, K.R. Khrapko, V.V. Shik, and A.D. Mirzabekov. (1988). "Determination of the Nucleotide Sequence of DNA Using Hybridization with Oligonucleotides. A New Method." *Doklady Akademii Nauk SSSR* 303, 1508–1511.

Markiewicz, W.T., K. Andrych-Rożek, M. Markiewicz, A. Żebrowska, and A. Astriab. (1994). "Synthesis of Oligonucleotides Permanently Linked with Solid Supports for Use as Synthetic Oligonucleotide Combinatorial Libraries. Innovations in Solid Phase Synthesis." In R. Epton (ed.), *Biological and Biomedical Applications*. Birmingham: Mayflower Worldwide, pp. 339–346.

Pevzner, P.A. (1989). "l-tuple DNA Sequencing: Computer Analysis." *Journal of Biomolecular Structure and Dynamics* 7, 63–73.

Southern, E.M. (1988). United Kingdom Patent Application GB8810400.

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