Tabu search based algorithms for DNA sequencing

Jacek Blazewicz Marta Kasprzak Aleksandra Swiercz

Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland, and Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12, 61-704 Poznan, Poland

{jblazewicz,mkasprzak,aswiercz}@cs.put.poznan.pl

1 DNA sequencing problem

The *DNA sequencing* is an elementary approach in computational molecular biology, leading to recognizing genetic information of organisms. The information is encoded as a sequence of nucleotides (basic particles of DNA) composing double helix, which in human beings reaches the length of 3 billions. There are four types of the nucleotides: A, C, G, and T (the abbreviations of names of their nitrogenous bases: adenine, cytosine, guanine, and thymine). The order of nucleotides in DNA strands determines processes occuring in organisms, thus their structures and functions. The sequencing is the process of reading the nucleotide order of an unknown DNA fragment, of length usually up to 1000 nucleotides (the length depends on kind of biological experiment). Such sequences are further combined into larger contigs in the assembly process and analyzed toward extraction of genetic information.

The approach to DNA sequencing discussed here is *DNA sequencing by hybridization* (SBH). It consists of two stages: first the biological data are produced in the hybridization experiment, next they become input to the computational phase, which ends with the reconstructed sequence. The output of the hybridization reaction can be viewed as a set (called *spectrum*) of words (*oligonucleotides*) over the alphabet {A, C, G, T}, being short subsequences of the studied DNA fragment. The aim of the *DNA sequencing problem* is to reconstruct the original DNA sequence of a known length on the basis of these overlapping oligonucleotides [20].

In the *standard approach* to SBH, the oligonucleotide library used in the hybridization experiment contains all possible oligonucleotides of a given constant length. The spectrum being output of the experiment is a subset of the library, i.e. the set of words of equal length composing the original sequence [22]. Contrary to it, in the *isothermic approach* to SBH, the library contains all oligonucleotides of constant temperature of melting oligonucleotide duplexes, but differing in lengths. This is due to assure more perfect chemical conditions, what results in lower number of experimental errors. In order to provide the spectrum with the certainty that the whole studied DNA fragment is covered by the oligonucleotides, the experimental phase must be carried out with two isothermic libraries differing by two degrees [1]. There are also another propositions modifying the standard SBH procedure, like for example multistage SBH [17] or SBH with universal nitrogenous bases [21].

For both standard and isothermic SBH, the computational complexity of several variants of the combinatorial problem is already known and the corresponding variants of the two approaches belong to the same complexity classes. The variants with no errors in the spectrum are polynomially solvable [19, 8] while the variants assuming presence of errors in the data (negative errors, positive errors, or both) are all strongly NP-hard [7, 8]. Because errors are present in outcomes of biological experiments as a rule, algorithms constructed for DNA sequencing problem should be heuristics.

2 Methods

Since the first formulation of the DNA sequencing problem with errors (as a variant of Selective Traveling Salesman Problem) and the first exact branch-and-bound algorithm presented in the same paper [3], next methods dedicated to SBH with errors were heuristics. A few constructive heuristics are known [19, 23, 14], but majority of published algorithms have been based on metaheuristic schemes: tabu search, evolutionary techniques [9, 13, 10, 12], ant colony optimization [11] and others [18].

Focusing our attention to tabu search, there are a few methods solving the problem in both standard and isothermic version [4, 5, 2]. Although the input data in standard and isothermic SBH differ (constant vs. variant lengths, one vs. two libraries), the general schemes of tabu search used for solving corresponding problems are similar. The goal is to maximize the number of elements from a spectrum composing a solution, which is a sequence of nucleotides of given length. The spectrum is represented in the algorithms by two data structures: an ordered list of oligonucleotides constituting a current solution, and an unordered set of remaining oligonucleotides. Three kinds of moves are defined on two kinds of objects: insertion into solution, deletion from solution, and shift within solution, with the objects being an oligonucleotide or a cluster (a group of strongly connected neighboring oligonucleotides in the solution). Several other rules, bounding the usage of the moves are defined, to reach the best local solution in reasonable time. Two criterion functions are used simultaneously: the condensation function (the ratio of the number of oligonucleotides in the solution to the solution length), which conduces to the creation of clusters, and the extending function (the number of oligonucleotides in the solution) lengthening the solution. The tabu list remembers inserted or shifted oligonucleotides.

The method from [5] additionally utilizes frequency-based memory and scatter search [16] (the latter in restarts of the search procedure) as parts of the diversification strategy. This combination of tabu and scatter search appeared to be very successful, with quality of results reaching 99.5% for the largest instances. Later this method has been superseded by the following proposals: the hybrid GA [6], ant colony [11], and constructive [14] heuristics.

For isothermic SBH the best algorithm so far is a hybrid GA [10, 6] that joins certain elements of a genetic algorithm with associated elements of a simple tabu search approach. The resulting method has additional search capabilities and performs significantly more effectively than the genetic algorithm component by itself. The algorithm also adopts a special rule for combining solutions proposed in connection with tabu search [15] which employs structured combination of vectors and voting evaluations. The individuals are treated as vectors to establish precedence relationships between oligonucleotides. The creation of the offspring requires at each step the selection of an oligonucleotide by choosing the best vote from two vectors (parents). Hence, the offspring inherits the best characteristics (votes) from parents (vectors). The resulting algorithm proves both to be robust and to yield solutions of particularly high quality, 100% in most of the cases.

References

[1] J. Blazewicz, P. Formanowicz, M. Kasprzak, and W.T. Markiewicz, Method of sequencing of nucleic acids, Polish patent application P335786 (1999).

[2] J. Blazewicz, P. Formanowicz, M. Kasprzak, W.T. Markiewicz, and A. Swiercz, Tabu search algorithm for DNA sequencing by hybridization with isothermic libraries, *Computational Biology and Chemistry* 28 (2004) 11–19.

[3] J. Blazewicz, P. Formanowicz, M. Kasprzak, W.T. Markiewicz, and J. Weglarz, DNA sequencing with positive and negative errors, *Journal of Computational Biology* 6 (1999) 113–123.

[4] J. Blazewicz, P. Formanowicz, M. Kasprzak, W.T. Markiewicz, and J. Weglarz, Tabu search for DNA sequencing with false negatives and false positives, *European Journal of Operational Research* 125 (2000) 257–265.

[5] J. Blazewicz, F. Glover, and M. Kasprzak, DNA sequencing — tabu and scatter search combined, *INFORMS Journal on Computing* 16 (2004) 232–240.

[6] J. Blazewicz, F. Glover, M. Kasprzak, W.T. Markiewicz, C. Oguz, D. Rebholz-Schuhmann, and A. Swiercz, Dealing with repetitions in sequencing by hybridization, *Computational Biology and Chemistry* 30 (2006) 313–320.

[7] J. Blazewicz and M. Kasprzak, Complexity of DNA sequencing by hybridization, *Theoretical Computer Science* 290 (2003) 1459–1473.

[8] J. Blazewicz and M. Kasprzak, Computational complexity of isothermic DNA sequencing by hybridization, *Discrete Applied Mathematics* 154 (2006) 718–729.

[9] J. Blazewicz, M. Kasprzak, and W. Kuroczycki, Hybrid genetic algorithm for DNA sequencing with errors, *Journal of Heuristics* 8 (2002) 495–502.

[10] J. Blazewicz, C. Oguz, A. Swiercz, and J. Weglarz, DNA sequencing by hybridization via genetic search, *Operations Research* 54 (2006) 1185–1192.

[11] C. Blum, M.Y. Valles, and M.J. Blesa, An ant colony optimization algorithm for DNA sequencing by hybridization, *Computers and Operations Research* 35 (2008) 3620–3635.

[12] C.A. Brizuela, L.C. Gonzalez-Gurrola, A. Tchernykh, and D. Trystram, Sequencing by hybridization: an enhanced crossover operator for a hybrid genetic algorithm, *Journal of Heuristics* 13 (2007) 209–225.

[13] T.N. Bui and W.A. Youssef, An enhanced genetic algorithm for DNA sequencing by hybridization with positive and negative errors, *Lecture Notes in Computer Science* 3103 (2004) 908-919.

[14] Y. Chen and J. Hu, eSBH: an accurate constructive heuristic algorithm for DNA sequencing by hybridization, *Proceedings of 10th IEEE International Conference on Bioinformatics and Bioengineering*, Philadelphia (2010).

[15] F. Glover, Tabu search for nonlinear and parametic optimization (with links to genetic algorithm), *Discrete Applied Mathematics* 49 (1994) 231-255.

[16] F. Glover, Scatter search and path relinking, *in*: D. Corne, M.Dorigo, F. Glover (Eds.) *New Ideas in Optimization*, McGraw-Hill, New York (1999) 297-316.

[17] S. Kruglyak, Multistage sequencing by hybridization, *Journal of Computational Biology* 5 (1998) 165-171.

[18] A. Nikolakopoulos and H. Sarimveis, A metaheuristic approach for the sequencing by hybridization problem with positive and negative errors, *Engineering Applications of Artificial Intelligence* 21 (2008) 247–258.

[19] P.A. Pevzner, l-tuple DNA sequencing: computer analysis, *Journal of Biomolecular Structure and Dynamics* 7 (1989) 63-73.

[20] P.A. Pevzner, *Computational Molecular Biology: An Algorithmic Approach*, MIT Press, Cambridge (2000).

[21] A.M. Frieze, F.P. Preparata, and E. Upfal, Optimal reconstruction of a sequence from its probes, *Journal of Computational Biology* 6 (1999) 361-368.

[22] E.M. Southern, United Kingdom patent application GB8810400 (1988).

[23] J.H. Zhang, L.Y. Wu, and X.S. Zhang, Reconstruction of DNA sequencing by hybridization, *Bioinformatics* 19 (2003) 14-21.