

A TSP FORMULATION OF THE DNA SEQUENCING PROBLEM

Athanassios Nikolakopoulos and Haralambos Sarimveis

Corresponding author: Haralambos Sarimveis, Assistant Professor at National Technical University of Athens, School of Chemical Engineering, e-mail: hsarimv@central.ntua.gr

Nikolakopoulos Athanassios, PhD student at National Technical University of Athens, School of Chemical Engineering, e-mail: thnikola@central.ntua.gr

National Technical University of Athens, School of Chemical Engineering,
9, Heron Polytechniou Str. Zografou Campus, Athens 15780, Greece

Key words: DNA sequencing with errors, Meta-heuristic Search, Traveling Salesman Problem, Threshold Accepting

ABSTRACT

The main target of this paper is the formulation of the Sequencing by Hybridization problem (*SbH*) (Fodor et al., 1991) as an Asymmetric Traveling Salesman problem (*ATSP*) (Reinelt, 1994). Furthermore, a metaheuristic method is proposed for the solution of the problem. *SbH* is a method where an original DNA string of up to 1000 nucleotides is hybridized with a bio-chip that contains all sequences of nucleotides of length l ($l \in \{8, 10, 12\}$). The nucleotides belong to the set: {adenine (a), cytosine (c), guanine (g) and thymine (t)}. Thus, the whole DNA chain can be represented as a string comprised of a sequence of the four letters (a , c , g and t) repeated in an arbitrary manner. The purpose of this bioinformatics method is to identify all the nucleotides and their related positions in the original string. All the identified sequences of length l , are called oligonucleotides and comprise the *spectrum*. Ideally, all the oligonucleotides are identical to fragments of the original sequence, overlap one over the immediately succeeding oligonucleotide by $l-1 = 2$ nucleotides, so they reach a total number of oligonucleotides: $n - l + 1$. Unfortunately realistic spectrums contain errors. An error is called *negative* if an oligonucleotide from the ideal spectrum is not contained in the experimental spectrum, and *positive* if the reverse case occurs (Blazewicz et al., 1999). For example we assume a spectrum that comprises three oligonucleotides **{1, 2, 3}** of length $l = 3$, where **1**:= { a, c, g }, **2**:= { c, g, t }, **3**:= { t, t, a } and an original DNA string:= { c, g, t, a, c, g }. Consider the arbitrary order of the oligonucleotides, *order* = [**2 3 1**]. The constructed DNA string is := { c, g, t, t, a, c, g } consists of 7 nucleotides because the 2nd overlaps the 3rd and the 3rd overlaps the 1st oligonucleotide by one nucleotide. It is obvious that the spectrum is not ideal since not all

neighboring oligonucleotides overlap by $l-1 = 2$ nucleotides. For example the 3rd oligonucleotide is overlapped by the 2nd by $k = 1$ (t) nucleotides and since $l-1 = 2$, $k < l-1$. An example of a positive error is that the oligonucleotide ($g t a$) is missing from the spectrum though it is contained in the original DNA sequence. An example of negative error is oligonucleotide 3:=($t t a$) which is included in the spectrum though it is not contained in the DNA sequence. Such a data base of oligonucleotides that possibly contain errors must be processed appropriately to formulate a sequence of nucleotides that matches as close as possible the original DNA string. The bigger is the overlapping between successive oligonucleotides the bigger becomes the resemblance of the constructed DNA string to the original.

In the present work the oligonucleotides are considered as nodes of an ATSP. Since the nucleotide denoting letters of any two oligonucleotides i, j are known the maximum overlaps:

$$Ov_{ij} \quad i = 1, \dots, m \quad \text{and} \quad j = 1, \dots, m \quad (m: \text{total number of oligonucleotides})$$

of the last letters of i over the first letters of j can be easily calculated (see oligonucleotides **2** and **3** from the last example $Ov_{12} = 2$ and $Ov_{21} = 0$). Knowing the overlaps we compute the distances between the oligonucleotides by the following equation:

$$d_{ij} = 2 \cdot l - Ov_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, m$$

In general $d_{ij} \neq d_{ji}$ (asymmetric problem $d_{12} = 4 \neq d_{21} = 6$). The proposed solution is an order of oligonucleotides $\mathbf{P} = \{P_1, P_2, \dots, P_m\}$. The objective function of the problem is the minimization of the sum of the distances of sequential nodes in the proposed solution:

$$\text{Obj:} \quad \min \sum d_{P_i, P_{i+1}}, \quad i = 1, \dots, m$$

For the computational experiments we use DNA data from GenBank and produce erroneous spectrums for original strings containing 100 to 1000 nucleotides. The percentages of negative or/and positive errors in the produced spectrums vary from 10 to 30%. Before sequencing the oligonucleotides of the *spectrum*, the possible positive errors are recognized according to their maximum overlap and removed. If the procedure does not reach a feasible solution, the data corresponding to the less possible positive error are reinserted in the *spectrum* and the problem is resolved again. For the sequencing problem a variation of the Threshold Accepting (TA) (Dueck and Scheuer, 1990) method with intense Local Search (Salvesberg, 1992, Cirasella et al. 2001, Helsgaun, 2000, Lin and Kernighan, 1973) is used. The goal is to minimize the total traveled distance of the ATSP route, thus attaining the tightest bonding of the oligonucleotides and therefore the best identification to the original string. The current solution of the TA method is a permutation of the available oligonucleotides and the search is advanced by applying local moves (such as the k -Opt moves - Lin and Kernighan, 1973) in order to produce a new permutation. For each permutation (say [**3 2 1**] for example 1), an *Insertion method* constructs progressively a sequence of oligonucleotides ([**1 2 3**] with DNA string $\{a, c, g, t, t, a\}$) by inserting them one by one in a position that corresponds to the minimal cost:

$$\underset{i}{\operatorname{argmin}}(d_{ik} + d_{kj} - d_{ij}) \quad \text{when the oligonucleotide } k \text{ is inserted after } i$$

The solution [**1 2 3**] for example 1 is not correct but acceptable because the length of the constructed DNA string is equal to the length of the original string. The solution that corresponds to each permutation is evaluated by calculating the sum of all distances between successive oligonucleotides. TAs' iterative procedure is controlled by a problem-size *diminishing shell*. The *diminishing shell* is a

procedure that interrupts the execution of the TA algorithm at appropriate time instances. This procedure accomplishes a reduction of the size of the problem by grouping the tightly sequenced oligonucleotides into clusters of oligonucleotides, and formulates a new *ATSP* of smaller dimension. The nodes of this smaller *ATSP* are the oligonucleotide clusters. The entire procedure involving both the size diminishing shell and the TA algorithm is repeated till an acceptable solution (the length of the constructed DNA string is equal to the length of the original DNA string) is reached. The proposed method provides solutions of better quality compared to algorithms that have been presented in the recent bibliography (Blazewicz et al., 2002, Endo, 2004). Moreover it produces solutions for problems of bigger size i.e. DNA strings of bigger cardinality or/and with larger percentage of erroneous data.

ACKNOWLEDGEMENTS

Financial support by FAGE S.A. and the General Secretariat of Research and Technology in Greece under the PENED 2001 research program (01EΔ38) is gratefully acknowledged.

REFERENCES

1. Fodor, S.P.A., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis, *Science* 251 767–773.
2. Reinelt, G. (1994). The Traveling Salesman: Computational Solutions for TSP Applications, *Lecture Notes in Computer Science*, Springer-Verlag, Vol 810.
3. Blazewicz, J., Formanowicz, P., Kasprzak, M., Markiewicz, W.T. and Weglarz, J. (1999). DNA sequencing with positive and negative errors. *Journal of Computational Biology*, 6, 113–123.
4. Dueck, G., Scheuer, T. (1990). Threshold accepting. A general purpose optimization algorithm appearing superior to simulated annealing, *Journal of Computational Physics*, 90, 161–175.
5. Savelsbergh, (1992). The vehicle routing problem with time windows: minimizing route duration, *ORSA Journal on Computing*, 4, 146–154.
6. Cirasella, J., Johnson, D. S., McGeoch, L. A. and Zhang, W. (2001). The asymmetric traveling salesman problem: Algorithms, instance generators, and tests. In Proc. 3rd ALENEX, *LNCS*, Springer-Verlag, 2153, 32-59.
7. Helsgaun, K. (2000). An effective implementation of the Lin-Kernighan traveling salesman heuristic, *European Journal of Operations Research*, 12, 106-130.
8. Lin, S., Kernighan, B.W. (1973). An Effective heuristic algorithm for the traveling salesman problem, *Operations Research*, 21, 972-989.
9. Blazewicz, J., Formanowicz, P., Guinand, F. and Kasprzak, M. (2002). A heuristic managing errors for DNA sequencing. *Journal of Bioinformatics*, 18(5), 652–660.
10. Endo, A. Takaho, (2004). Probabilistic Nucleotide Assembling Method for Sequencing by Hybridization, *Bioinformatics*, 20(14), 2181-2188.