Distance-Constrained Elementary Path Problem : New MIP Formulations

Stefan Balev[1], Rumen Andonov[2], Nicola Yanev[3]

[1] Normandie Univ, UNIHAVRE, LITIS, Le Havre, France stefan.balev@univ-lehavre.fr
[2] Univ Rennes, Inria, CNRS, IRISA, Rennes, France, randonov@irisa.fr
[3] Department of Informatics, Center for Advanced Bioinformatics Research, South-West University Neofit Rilski, Bulgaria choby@math.bas.bg

Mots-clés : *Graphs, Longest Path Problem, De novo Genome assembly*

1 Introduction

The *Distance-Constrained Elementary Path (DCEP)* problem can be formulated as follows : we are given a directed graph $G = (V, E, l)$, where $l_e \geq 0$ denotes the length associated with each arc $e \in E$. In addition, it is given a set DC of vertex pairs (called *distance constraints*) such that any couple $(u, v) \in DC$ is associated with a distance interval $dc(u, v) = [\underline{d}(u, v), \overline{d}(u, v)]$. We say that a path $\widetilde{\mathcal{P}}$ in G satisfies the distance constraint $dc(u, v)$ if both u and v are on $\widetilde{\mathcal{P}}$ and the subpath of $\widetilde{\mathcal{P}}$ between u and v has length in $dc(u, v)$. The goal is to find an elementary path in G that maximizes the number of satisfied distance constraints. DCEP has been introduced in one of our previous papers where two MIP formulations have been also given [8]. Here we introduce a new formulation and continue the analysis of these formulations.

2 DCEP motivation and similarity with famous problems

DCEP is motivated by the genome assembly problem which is a challenging computational task in bioinformatics aiming at reconstructing the full genome of an organism from short DNA sequences (*reads*) [9]. No satisfactory solution for DCEP is known today and heuristics are essentially described in the literature [3, 4, 11]. The methodology that we propose in [6, 7] differs significantly from these heuristics since it is based on integer programming approach for solving genome assembly as a problem of finding an elementary path in a specific graph, which satisfies additional constraints encoding the insert-size (distance) information. We thus develop a global optimization approach where the various assembly steps are simultaneously solved in the framework of a common objective function. The numerical experiments show that our tool produces assemblies of significantly higher quality than some widely-used heuristics on a chloroplast genomes benchmark [1]. These results justify the efforts for designing exact approaches for genome assembly.

3 Our approach

While the paper [7] is application oriented, in [8] we revisit the mixed integer linear programming formulation proposed there from a combinatorial optimization viewpoint. Furthermore, we show how to adapt the well known Miller, Tucker and Zemlin (MTZ) formulation for solving the longest path problem [10] in order to solve the DCEP problem. Note that the challenges in DCEP problem are somehow similar, but harder in practice, to the ones in solving longest elementary path problem [5, 12]. For example one of the hardest constraints in the MIP formulation for DCEP is the sub-tour elimination contraint.

Here we propose a new MIP formulation for DCEP which is inspired by the so called "alignment graph formulation" described in [2, 13]. We adapt the later one to the specificities of DCEP by introducing the above mentioned distances contraints. We furthermore perform comparisons with the formulations described in [8] on a set of simulated and real instances.

4 Conclusions

While our exact approach based on MIP formulation for the DCEP problem clearly outperforms the heuristics approaches in terms of the quality of the results, we are currently unable to assemble huge genomes in a reasonable time. Accelerating resolving the DCEP problem without sacrificing the quality of the obtained results is the main focus of our current research.

Références

[1] Rumen Andonov, Hristo Djidjev, Sébastien François, and Dominique Lavenier. Complete Assembly of Circular and Chloroplast Genomes Based on Global Optimization. *Journal of Bioinformatics and Computational Biology*, pages 1–28, 2019.

[2] Stefan Balev. Solving the protein threading problem by lagrangian relaxation. In Inge Jonassen and Junhyong Kim, editors, *Algorithms in Bioinformatics*, pages 182–193, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[3] A. Bankevich, S. Nurk, D. Antipov, A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner. Spades : a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, 19(5) :455–477, May 2012.

[4] Marten Boetzer, Christiaan V. Henkel, Hans J. Jansen, Derek Butler, and Walter Pirovano. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics (Oxford, England)*, 27(4) :578–579, February 2011.

[5] Q. T. Bui, Y.c Deville, and Q. D. Pham. Exact methods for solving the elementary shortest and longest path problems. *Annals of Operations Research*, 244(2) :313–348, 2016.

[6] S. François, R. Andonov, H. Djidjev, and D. Lavenier. Global optimization methods for genome scaffolding. *Electronic Notes in Discrete Mathematics*, 64, 2018.

[7] S. François, R. Andonov, D. Lavenier, and H. Djidjev. Global optimization approach for circular and chloroplast genome assemblies. In *10th International Conference on Bioinformatics and Computational Biology (BICoB 2018)*, 2018. Mars, 2018, Las Vegas, USA.

[8] Sébastien Francois, Rumen Andonov, Hristo Djidjev, Metodi Traikov, and Nicola Yanev. Mixed Integer Linear Programming Approach for a Distance-Constrained Elementary Path Problem. In *16th Cologne-Twente Workshop on Graphs and Combinatorial Optimization*, Paris, France, June 2018.

[9] Daniel H. Huson, Knut Reinert, and Eugene W. Myers. The greedy path-merging algorithm for contig scaffolding. *J. ACM*, 49(5) :603–615, 2002.

[10] C. E. Miller, A. W. Tucker, and R. A. Zemlin. Integer programming formulation of traveling salesman problems. *J. ACM*, 7(4) :326–329, October 1960.

[11] K. Sahlin, F. Vezzi, B. Nystedt, J. Lundeberg, and L. Arvestad. BESST - efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics*, 15 :281, 2014.

[12] Leonardo Taccari. Integer programming formulations for the elementary shortest path problem. *European Journal of Operational Research*, 252(1) :122–130, 2016.

[13] Nicola Yanev, Rumen Andonov, Philippe Veber, and Stefan Balev. Lagrangian approaches for a class of matching problems in computational biology. *Computers & Mathematics with Applications*, 55(5) :1054 – 1067, 2008. Modeling and Computational Methods in Genomic Sciences.