Theoretical Computer Science 595 (2015) 92–106

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

A complexity and approximation framework for the maximization scaffolding problem

A. Chateau [a,b,*], R. Giroudeau [a,*]

[a] LIRMM – Université de Montpellier – UMR 5506 CNRS, 161 rue Ada, 34392 Montpellier Cedex 5, France
[b] Institut de Biologie Computationnelle, LIRMM Bât. 5 – 860 rue de St Priest, 34090, Montpellier, France

ARTICLE INFO

Article history:
Received 28 July 2014
Received in revised form 19 May 2015
Accepted 7 June 2015
Available online 11 June 2015
Communicated by A. Marchetti-Spaccamela

Keywords:
Complexity
Polynomial-time approximation algorithm
Scaffolding

ABSTRACT

We explore in this paper some complexity issues inspired by the contig scaffolding problem in bioinformatics. We focus on the following problem: given an undirected graph with no loop, and a perfect matching on this graph, find a set of cycles and paths covering every vertex of the graph, with edges alternatively in the matching and outside the matching, and satisfying a given constraint on the numbers of cycles and paths. We show that this problem is $\mathcal{NP}$-complete, even in planar bipartite graphs. Moreover, we show that there is no subexponential-time algorithm for several related problems unless the Exponential-Time Hypothesis fails. Lastly, we also design two polynomial-time approximation algorithms for complete graphs.

1. Introduction

We investigate the complexity of a problem inspired by the scaffolding problem, coming from bioinformatics. This problem concerns one of the computational steps involved in the production of the whole DNA sequence of a new genome. Indeed, it is not possible, due to technological issues, to read the whole sequence directly from the DNA molecule. Instead, the sequence is built through different steps, each of them presenting algorithmic challenges. Some approximation algorithms have been proposed around the assembly problem modelized as a Shortest Superstring Problem [26]. This problem consists in finding the best way to assemble DNA fragments into longer sequences, by optimizing their overlap. We focus here on the contig scaffolding step, which consists, given a set of sequences of various lengths called *contigs*, produced by the assembly step, to infer the order and the orientation of the contigs along the genome, using a set of possibly inconsistent pairing information. This step produces *scaffolds*, which are lists of oriented contigs. First described in [14], it was presented as a problem of path merging in a particular kind of graphs, and was stated as $\mathcal{NP}$-complete. Further studies describe different types of heuristics and computational approaches [6,7,12], or compare the accuracy of the exact approach to the heuristics [9], but none of them further investigates the complexity aspects of the problem, especially in terms of approximation. We focus here on this latter aspect.

A more general framework was recently proposed in [4], where the problem is presented as the resolution of consecutive ones property problem on matrices encoding hypergraphs. The authors include multiplicity ranges on the contigs, meaning that a contig may be repeated in the scaffolds, and propose interesting approximation results in the case where there is no constraint on the number of paths and cycles. This approach can use phylogenetic information, instead of the classical

* Corresponding authors.
E-mail addresses: chateau@lirmm.fr (A. Chateau), rgirou@lirmm.fr (R. Giroudeau).

http://dx.doi.org/10.1016/j.tcs.2015.06.023
0304-3975/

Table 1
Synthetic complexity results for Scaffold problems.

Problem	Complexity
(σ_p, σ_c)-SP with $n = \sigma_p + 2\sigma_c$	$\mathcal{P}$ (Theorem 1)
(σ_p, σ_c)-SP	$\mathcal{NP}$-Complete (Theorem 2)
(σ_p, σ_c)-SP in planar bipartite graphs	$\mathcal{NP}$-Complete (Theorem 2)
$((\sigma_p, l_p \geq 2), (\sigma_c, 6))$-SP	$\mathcal{NP}$-Complete (Theorem 3)
$((\sigma_p, l_p \geq 2), (\sigma_c, 12))$-SP in bipartite graphs	$\mathcal{NP}$-Complete (Theorem 3)

Table 2
Synthetic approximation and non-approximability results for Scaffold problems.

Problem	Ratio	Complexity (time)
(σ_p, σ_c)-SP	$\rho \leq 3$ (Theorem 10 & Theorem 7)	$\mathcal{O}(n^2 \log n)$ (Algorithm 1) $\mathcal{O}(n^3)$ (Algorithm 4)
$(0, 1)$-SP	$\rho \leq 2$ (Theorem 11)	$\mathcal{O}(n^3)$ (Algorithm 4)
Min-CL-(σ_p, σ_c)-SP	$\rho \geq 7/6$ (Corollary 5)	–
Min-CL-(σ_p, σ_c)-SP bipartite g.	$\rho \geq 13/12$ (Corollary 5)	–

Table 3
Lower bounds for exact exponential-time algorithms for Scaffold problems.

Problem	Lower bound for exact Time-complexity
(σ_p, σ_c)-SP in bipartite graphs	$\mathcal{O}(2^{o(n)})$ (Corollary 6)
(σ_p, σ_c)-SP in planar bipartite graphs	$\mathcal{O}(2^{o(\sqrt{n})})$ (Corollary 6)
Min-CL-(σ_p, σ_c)-SP with degree ≤ 9	$\mathcal{O}(2^{o(n)})$ (Corollary 6)
Min-CL-(σ_p, σ_c)-SP with degree ≤ 5 bipartite graphs	$\mathcal{O}(2^{o(\sqrt{n})})$ (Corollary 6)

use of paired fragments. Other approaches use this kind of phylogenetic information [13,17]. Also recently, several types of information have been mixed to infer scaffolds from the assembly data, for instance in [1], where the chromatin structure of the chromosomes has been added to the classical mate-pair information to complete the human, mouse and drosophila genomes. This general trend highlights the need of flexible tools to efficiently solve 'scaffolding-like' problems.

In [2], we presented an alternative formalization of the problem, inspired by a variant of the very well known Traveling Salesman Problem. We also proved some preliminary results concerning complexity and approximation. This model is more general than the one proposed in [14], and allows to integrate the desired structure of the genome (number of circular or linear chromosomes). In a nutshell, the problem consists in finding in a graph called the scaffold graph, a disjoint cover by a fixed number of paths and cycles, of optimal total weight.

A very huge literature has been provided concerning the Traveling Salesman Problem and its variants. We refer the reader to [22] for an overview on the domain. Concerning the more general problem of finding a cover with a fixed number of vertex disjoint cycles and paths, the papers especially focus on feasibility criteria, like sufficient conditions, typically on the degrees of the vertices, for the graph to admit such a cover (see, for instance, [5]). The cases where the numbers of paths and cycles are not fixed define a wide range of possibilities. Indeed, finding an optimal cover by disjoint cycles is a polynomial problem, when the number of cycles is not fixed [24]. On the contrary, finding an optimal cover by disjoint paths with at least two edges is $\mathcal{NP}$-complete [23]. Also, the problem to infer the cycles-number of a cycle cover, known as cycle packing, has been already studied: finding the minimal number of cycles which are necessary to cover a graph is $\mathcal{NP}$-complete [18]. Anyway, concerning the problem of finding, and optimizing, a spanning subgraph with a fixed number of cycles and paths, this is to our knowledge the first study of this kind of problem in terms of complexity and approximability.

In this article, we extend the results stated in [2] in several directions. We explore a polynomial case, we prove some lower bounds on polynomial-time approximation and on subexponential-time algorithms in the case where the aim is the minimization of the length of the cycles, and describe polynomial-time approximation algorithms for the case of the maximization of the weight of the solution is considered. The results are summarized in Tables 1, 2 and 3.

This article is organized as follows: the next section is devoted to formal description of the Scaffold problems. In Sections 3 and 4 we pay attention to computational complexity and several non-approximability results whereas in Section 5 we design some polynomial-time approximation algorithms for the maximization problem.

2. Formal description of the problems

In what follows, we consider $G = (V, E)$ an undirected graph with an even number $2n$ of vertices and without self loops. We suppose that there exists a perfect matching in G, denoted by M^*. Let $w : E \to \mathbb{N}$ be a weight function on the edges. In the bioinformatic context, edges in M^* represent contigs, and the other edges figure the ways to link the contigs together, their weight representing the support of each of these hypotheses (e.g. the number of pairs of reads matching on both contigs).