Available online at www.sciencedirect.com

SCIENCE DIRECT

ELSEVIER

Discrete Applied Mathematics 154 (2006) 1709 – 1721

DISCRETE APPLIED MATHEMATICS

www.elsevier.com/locate/dam

Fast algorithms for identifying maximal common connected sets of interval graphs

Fabien Coulon[a], Mathieu Raffinot[b,*]

[a] *Laboratoire d'Informatique Fondamentale et Appliquée de Rouen (LIFAR), Faculté des Sciences, place Emile Blondel, 76821 Mont-Saint-Aignan, France*

[b] *CNRS-Poncelet Laboratory, Independent University of Moscow, 11 street, Bolchoï Vlassievski, 119 002 Moscow, Russia*

Received 19 May 2004; received in revised form 23 December 2005; accepted 15 February 2006
Available online 19 April 2006

Abstract

Given a family of interval graphs $F = \{G_1 = (V, E_1), \ldots, G_k = (V, E_k)\}$ on the same vertices V, a set $S \subset V$ is a maximal common connected set of F if the subgraphs of G_i, $1 \leqslant i \leqslant k$, induced by S are connected in all G_i and S is maximal for the inclusion order. The maximal general common connected set for interval graphs problem (gen-CCPI) consists in efficiently computing the partition of V in maximal common connected sets of F. This problem has many practical applications, notably in computational biology. Let $n = |V|$ and $m = \sum_{i=1}^{k} |E_i|$. For $k \geqslant 2$, an algorithm in $O((kn + m)\log n)$ time is presented in Habib et al. [Maximal common connected sets of interval graphs, in: Combinatorial Pattern Matching (CPM), Lecture Notes in Computer Science, vol. 3109, Springer, Berlin, 2004, pp. 359–372]. In this paper, we improve this bound to $O(kn \log n + m)$. Moreover, if the interval graphs are given as k sets of n intervals, which is often the case in bioinformatics, we present a simple $O(kn \log^2 n)$ time algorithm.

Keywords: Comparative genomics; Interval graph; Common connected set; Umbrella-free ordering

1. Introduction

Let $G = (V, E)$ be a loopless undirected graph. The degree of a vertex $x \in V$ in the graph G is denoted by $d_G(x)$. Let X be a subset of vertices of G, we denote $G[X]$ the subgraph induced by X: the set of vertices of $G[X]$ is X and its edge set is $E_X = E \cap \{(u, v) \mid u \in X, v \in X\}$.

A *connected component* of $G = (V, E)$ is a set $S \subset V$ that is connected and that cannot be augmented with other vertices.

Let F be a family of graphs on (or restricted to) the same vertices, say $F = \{G_1 = (V, E_1), \ldots, G_k = (V, E_k)\}$. We denote $n = |V|$ and $m = \sum_{i=1}^{k} |E_i|$. A connected set $X \subset V$ of F is such that each $G_i[X]$ is connected.

Definition 1. A set $S \subseteq V$ of vertices is a *maximal common connected set* of a family $F = \{G_1 = (V, E_1), \ldots, G_k = (V, E_k)\}$ of graphs if S is a maximal, with respect to the inclusion order, connected set of F.

[*] Corresponding author. Fax: +7 095 2916501.
E-mail address: mathieu@raffinot.net (M. Raffinot).

0166-218X/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2006.02.005

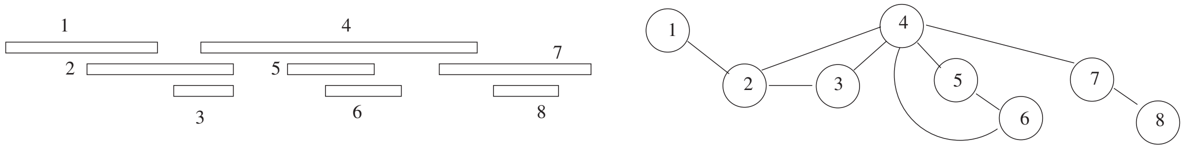

Fig. 1. A set of intervals and its corresponding interval graph.

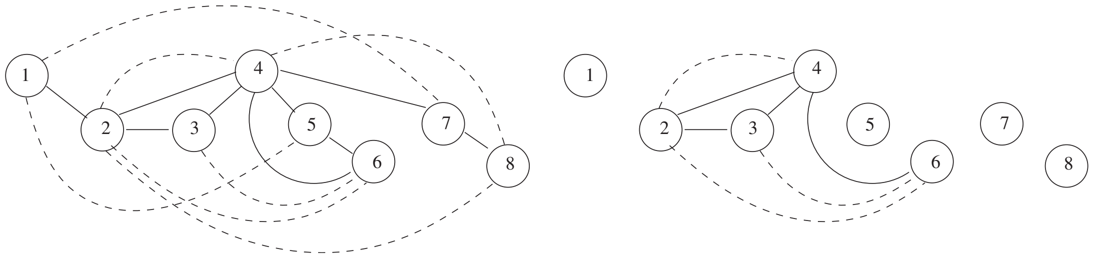

Fig. 2. Two interval graph families on the same vertices and the corresponding maximal common connected sets.

The *maximal common connected sets* of V obviously form a partition of V and the *general common connected problem* (gen-CCP), defined in [4], is to efficiently compute this partition. This problem arises in comparative genomics for the identification of clusters of genes/proteins/domains that are closely placed on chromosomes of several species, considering specific distances that take into account biological properties. The notions of *common intervals* [14,8] and *gene teams* [1] are two specific solutions designed for simple distances induced by positions on a linear model of chromosomes. The resulting clusters can be identified, respectively, in $\mathrm{O}(kn)$ and $\mathrm{O}(kn \log^2 n)$ worst case time. Gen-CPP solves the problem for three-dimensional distances in $\mathrm{O}(n \log n + m \log^2 n)$ time [4]. However, many of the distances appearing in computational biology are in fact given by interval graphs. This is mainly due to the fact that large biological contigs are built through interval graphs of smaller sequences (cDNA, ESTs, etc.). Solving gen-CCP efficiently on interval graphs is a real challenge, that has already been addressed in [7]. This restriction is called general common connected problem on intervals (gen-CCPI).

Formally, a graph is an *interval graph* iff there is a one-to-one mapping between its vertices and a set of intervals on the real line such that two vertices are adjacent iff their corresponding intervals intersect [10]. Fig. 1 gives an example of such a graph. Let $F = \{G_1 = (V, E_1), \ldots, G_k = (V, E_k)\}$ be a family of interval graphs. Fig. 2 shows an example of such a family and the partition of its vertices into maximal common connected sets. If the family F only contains a single graph, the problem is reduced to searching for connected components and is efficiently solved in $\mathrm{O}(n+m)$ time. Otherwise, gen-CCPI is solved in [7] in $\mathrm{O}((kn + m) \log n)$ time. The algorithm maintains a dynamic representation of connected components for all the graphs of the family using forests of maximal clique paths. This representation is then combined with an Hopcroft-like partitioning framework [9,11,2,5,12].

In this paper we propose a new $\mathrm{O}(kn \log n + m)$ worst case time algorithm which is, in any case, faster than that of [7]. The algorithm uses a new dynamic representation of the connected components of an interval graph that is combined with a highly simplified Hopcroft like partitioning framework. The representation is based on a specific vertex ordering that verifies the umbrella-free property [13,3]. Compared to the algorithm of [7], the partitioning framework is very similar but simpler. It resembles the original gene teams identification algorithm [1]. However, our new representation is more difficult to manage and the partitioning operation is more delicate.

Moreover, as an interval graph represents a set of intervals on the real line, it may therefore be given as a set of n intervals instead of as a full interval graph (see Fig. 1). This is often the case in computational biology. Building the corresponding graph is $\mathrm{O}(n + m)$, and gen-CCPI may also be solved for a family of k sets of n intervals with our new algorithm in $\mathrm{O}(kn \log n + m)$ time. However, in this specific case, we exhibit a simple algorithm solving gen-CCPI in $kn \log^2 n$, independently of m. As m may be counted (without building the real graph) on n intervals in $\mathrm{O}(n)$, the algorithm is of use as soon as $m = \Omega(n \log^2 n)$.