Discrete Applied Mathematics 179 (2014) 129–138

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Computing the blocks of a quasi-median graph

Sven Herrmann *, Vincent Moulton

School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK

ARTICLE INFO

Article history:
Received 25 July 2012
Received in revised form 4 July 2014
Accepted 20 July 2014
Available online 8 August 2014

Keywords:
Quasi-median graph
Median graph
Most parsimonious trees
Steiner trees
Mitochondrial evolution

ABSTRACT

Quasi-median graphs are a tool commonly used by evolutionary biologists to visualise the evolution of molecular sequences. As with any graph, a quasi-median graph can contain cut vertices, that is, vertices whose removal disconnect the graph. These vertices induce a decomposition of the graph into blocks, that is, maximal subgraphs which do not contain any cut vertices. Here we show that the special structure of quasi-median graphs can be used to compute their blocks without having to compute the whole graph. In particular we present an algorithm that, for a collection of n aligned sequences of length m, can compute the blocks of the associated quasi-median graph together with the information required to correctly connect these blocks together in run time $\mathcal{O}(n^2m^2)$, independent of the size of the sequence alphabet. Our primary motivation for presenting this algorithm is the fact that the quasi-median graph associated to a sequence alignment must contain all most parsimonious trees for the alignment, and therefore precomputing the blocks of the graph has the potential to help speed up *any* method for computing such trees.

1. Introduction

Quasi-median graphs are a tool commonly used by evolutionary biologists to visualise the evolution of molecular sequences, especially mitochondrial sequences (Schwarz and Dür [19]; Ayling and Brown [1]; Bandelt et al. [5]; Huson et al. [15, Chapter 9]). They were introduced by Mulder [18, Chapter 6] and their application to molecular sequence analysis was introduced for binary sequences in (Bandelt et al. [5]) and for arbitrary sequences in (Bandelt et al. [4]). A quasi-median graph can be constructed for an alignment of sequences over any alphabet (Bandelt and Dür [3]); for binary sequences they are also known as *median graphs* (Bandelt et al. [5]). An example of a quasi-median graph associated to the hypothetical alignment of sequences s_1–s_9 is presented in Fig. 1.1 (see Bandelt and Dür [3] for more details on how to construct such graphs).

Here we are interested in computing the cut vertices of a quasi-median graph as well as an associated decomposition of the graph. Recall that given a connected graph $G = (V(G), E(G))$, consisting of a set $V = V(G)$ of vertices and a set $E = E(G)$ of edges, a vertex $v \in V$ is called a *cut vertex* of G if the graph obtained by deleting v and all edges in E containing v from G is disconnected (for the basic concepts in graph theory that we use see, for example, (Diestel [9])). For example, in the quasi-median graph in Fig. 1.1 the cut vertices are precisely the white vertices and the black vertex s_8. As with any graph, the cut vertices of a quasi-median graph decompose it into *blocks*, that is, maximal subgraphs which do not contain any cut vertices themselves. These blocks in turn, together with the information on how they are linked together, give rise to the *block decomposition* of the graph (see Section 5 for a formal definition of this decomposition that we shall use which is

* Corresponding author.
E-mail addresses: sherrmann@mathematik.tu-darmstadt.de (S. Herrmann), Vincent.moulton@cmp.uea.ac.uk (V. Moulton).

http://dx.doi.org/10.1016/j.dam.2014.07.013
0166-218X/

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	G	T	A	T	C	A	G	T	A	T	A	T
s_2	G	T	G	T	C	A	G	T	A	C	G	T
s_3	A	T	G	T	C	A	A	C	G	C	A	T
s_4	A	T	G	T	C	A	C	C	A	C	A	C
s_5	A	C	A	C	T	C	G	C	A	C	A	T
s_6	A	C	A	C	T	G	G	C	A	C	A	T
s_7	G	C	G	T	C	A	G	C	A	C	A	T
s_8	G	T	G	T	C	A	G	T	A	C	A	T
s_9	A	T	G	T	C	A	C	C	A	C	A	G
s_{10}	A	C	A	C	C	G	G	C	A	C	A	T

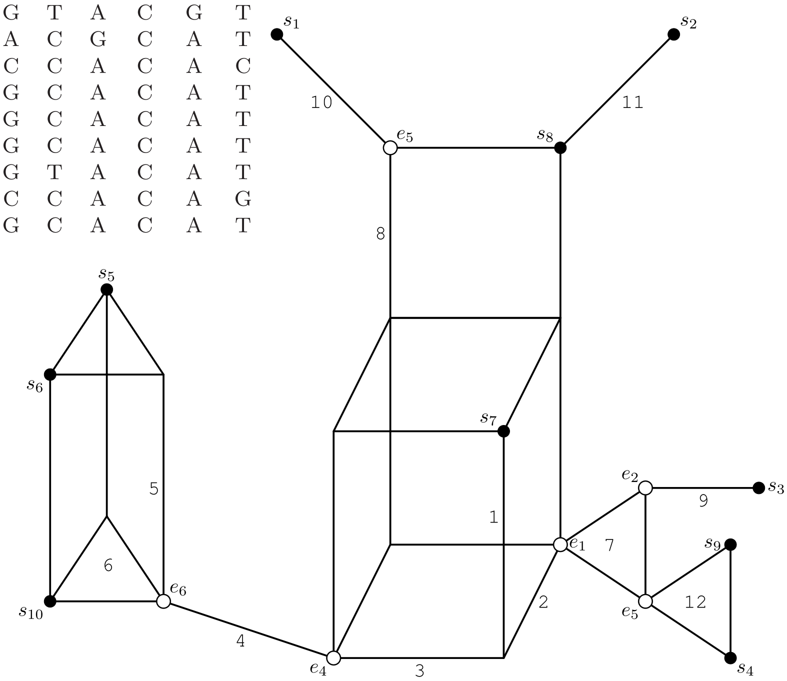

Fig. 1.1. An alignment of hypothetical DNA sequences and the associated quasi-median graph. The sequences correspond to the black vertices and the columns correspond to the sets of edges, as indicated by the labels.

specific to quasi-median graphs). It is well known that the block decomposition of a given graph can be computed in linear time from its vertices and edges; however, the size of a quasi-median graph is usually exponential in the size of the sequence alignment. Therefore, the main purpose of this paper is to provide an algorithm for computing the block decomposition of a quasi-median graph *without* having to compute the whole graph.

The results in this paper complement the well-developed theory of quasi-median networks (cf., e.g., (Bandelt et al. [7]; Imrich and Klavžar [16])). However, our primary motivation for computing the block decomposition of quasi-median graphs is provided by their close connection with most parsimonious trees (see, e.g., Felsenstein [13] for an overview of parsimony). Indeed, Bandelt and Röhl [8] showed that the set of *all* most parsimonious trees for a collection of (aligned, gap-free) sequences must be contained in the quasi-median graph of the sequences (see also (Bandelt [2]) for a proof of this result for median networks). More specifically, they showed that the most parsimonious trees for the sequences are in one-to-one correspondence with the Steiner trees for the sequences considered as a subset of the vertices of the quasi-median graph. It easily follows that the block decomposition of a quasi-median graph can be used to break up the computation of most parsimonious trees into subcomputations on the blocks. Of course, the quasi-median graph of an arbitrary collection of sequences may not contain any cut vertices but, as computing most parsimonious trees is NP-hard (Foulds and Graham [14]), it could still be a useful pre-processing step to compute the cut vertices of quasi-median graphs before trying to compute most parsimonious trees. Similarly, Misra et al. [17] propose an integer linear programme for computing a most parsimonious tree, which is based on the structure of the quasi-median graph (called the *generalised Buneman graph* by the authors). A computation of the block decomposition could be used to decompose the problem into smaller subproblems.

We now summarise the contents of the rest of this paper. We begin by presenting some preliminaries concerning quasi-median graphs in the next section. Then, in Section 3, we recall a characterisation of the vertices of a quasi-median graph given in (Bandelt et al. [6]), which we use in Section 4 to prove a key structural result for quasi-median graphs (Theorem 4.1). This result is a direct generalisation of Theorem 1 of (Dress et al. [6]) for median graphs and states that the blocks in a quasi-median graph are in bijection with the connected components of a certain graph which can be associated to an alignment that captures the degree of "incompatibility" between its columns. Using this result, we also derive a characterisation of the cut vertices of a quasi-median graph (Theorem 4.6). After defining the block decomposition of a quasi-median graph in Section 5, we present our algorithm for its computation in Section 6 (Algorithm 1). In particular, we prove that this algorithm correctly computes the block decomposition (Theorem 6.1) and also show that, for a collection of n aligned sequences of length m, the algorithm's run time is $\mathcal{O}(n^2m^2)$, independent of the size of the sequence alphabet (Theorem 6.3). We have implemented the algorithm and it is available for download at http://www.uea.ac.uk/computing/quasidec.