Discrete Applied Mathematics 197 (2015) 27–41

Contents lists available at ScienceDirect

Discrete Applied Mathematics

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

Discretization vertex orders in distance geometry

Andrea Cassioli [a], Oktay Günlük [b], Carlile Lavor [c], Leo Liberti [b,d,*]

[a] Mosek ApS, Fruebjergvej 3, 2100, Copenhagen, Denmark
[b] IBM Research, Yorktown Heights, NY 10598, USA
[c] Department of Applied Math. (IMECC-UNICAMP), University of Campinas, 13081-970, Campinas – SP, Brazil
[d] CNRS LIX, Ecole Polytechnique, 91128 Palaiseau, France

ARTICLE INFO

Article history:
Received 16 December 2013
Received in revised form 23 August 2014
Accepted 29 August 2014
Available online 23 September 2014

Keywords:
Molecular conformation
Protein
DDGP
DMDGP
Re-order
Branch-and-Prune

ABSTRACT

When a weighted graph is an instance of the Distance Geometry Problem (DGP), certain types of vertex orders (called discretization orders) allow the use of a very efficient, precise and robust discrete search algorithm (called Branch-and-Prune). Accordingly, finding such orders is critically important in order to solve DGPs in practice. We discuss three types of discretization orders, the complexity of determining their existence in a given graph, and the inclusion relations between the three order existence problems. We also give three mathematical programming formulations of some of these ordering problems.

1. Introduction

The Distance Geometry Problem (DGP) is as follows: given a positive integer K and a simple, undirected, nonnegatively weighted graph $G = (V, E, d)$, where $d : E \to \mathbb{R}_+$, find a realization $x : V \to \mathbb{R}^K$ such that:

$$\forall \{u, v\} \in E \quad \|x_u - x_v\|_2 = d_{uv}. \tag{1}$$

If G is disconnected then realizing G is the same as realizing its connected components, so we assume G is connected.

Solution methods for the DGP generally involve a search in continuous space [15]. On the other hand, several applications of the DGP supply some guarantees on the sparsity structure of the input graph, which may in turn imply graph rigidity. The solution set is then a finite subset of a Euclidean space, which allows for remarkable performance improvements of the solution algorithms. Many methods are iterative in nature: they assume a small subset of vertices have known positions, and try and infer the position of the rest of the vertices in some order. Thus, vertex orders play an important role. Trilateration orders, for example, guarantee that every vertex beyond the first $K + 1$ is adjacent to at least $K + 1$ predecessors [4]. This makes it possible to uniquely triangulate the position of each next vertex. This implies a polynomial time algorithm and a unique solution modulo translations and rotations.

The main focus of this paper is to determine the worst-case complexity class of many vertex ordering problems used in algorithms for solving the DGP on certain rigid graphs. We also propose and test three Mixed-Integer Linear Programming (MILP) formulations for solving such vertex ordering problems, and empirically determine that they can only be useful for rather small-scale instances.

* Corresponding author at: CNRS LIX, Ecole Polytechnique, 91128 Palaiseau, France. Tel.: +33 169334138; fax: +33 169334073.
E-mail addresses: andrea.cassioli@mosek.com (A. Cassioli), gunluk@us.ibm.com (O. Günlük), clavor@ime.unicamp.br (C. Lavor), leoliberti@us.ibm.com, liberti@lix.polytechnique.fr (L. Liberti).

http://dx.doi.org/10.1016/j.dam.2014.08.035
0166-218X/

1.1. Vertex orders in protein conformation

The function of proteins is strongly related to their chemical composition and their three-dimensional structure: proteins usually fold in space until they reach a stable configuration having low potential energy. Finding their 3D structure is therefore an important task in pharmaceutical research. Many approaches exist [26]: in this paper we adopt the point of view of distance geometry [1,2,23].

We represent a protein by means of a graph where vertices represent atoms and edges are present if the distance between two adjacent atoms is known. Atomic distances may be known for chemical or physical reasons, or because they were estimated using Nuclear Magnetic Resonance [27]. Proteins consists of chains of amino acids, which come in twenty different types. Amino acids consist of a common structure: a small chain starting with the amino group H_3N, followed by the α carbon C_α (linked to a hydrogen atom and to a side chain), followed by the carboxyl group COO^-. Amino acids only differ because of their side chains. The whole protein can then be seen as a backbone consisting of a chain formed by the repeated common structures of each amino acid, and many dangling side chains. To a large extent, the problem of finding a 3D realization of the protein can be decomposed into the subproblems of realizing the backbone and, separately, the side-chains; and then combining the partial realizations in a consistent way [25].

Protein backbones enforce an order on the atoms in the backbone. This order has some interesting properties: we know the distance of each atom v to its predecessor $v-1$, since covalent bond lengths are known for chemical reasons. Since covalent bond angles are also known, for every triangle of three consecutive atoms we know two of its side lengths and the angle between them: thus we can also compute the length of its third side, i.e. the distance between v and $v-2$. Moreover, NMR can estimate all distances up to a certain threshold (around 5.5 Å). It is known that distances between atom v and $v-3$ are always below this threshold, so the distance between v and $v-3$ is also known. This order makes the protein graph look like a chain of embedded cliques of size 4 (realized as 3-simplices), each sharing a face with the preceding one, plus possibly other edges called *pruning edges*. Pruning edges are due to the fact that when a protein backbone folds in space, two atoms might come to be physically close even though they are be very distant in terms of their ranks in the backbone order. In particular, their Euclidean distance becomes known when it is below the NMR threshold. Orders were each vertex is adjacent to three predecessors have been shown to yield rigid structures in early 1900s [7].

1.2. The Branch-and-Prune algorithm

This order was instrumental in devising a discrete method called Branch-and-Prune (BP) for finding the 3D realization of protein backbones [14]. Although the BP was not the first discrete method for this problem [3], it was the first which could find all incongruent solutions to any given problem instance. Most methods previously proposed in the literature, by contrast, were searches in continuous space (see [17] and references therein).

The principle behind the BP is that any 3D simplex on the vertices $\{v, v-1, \ldots, v-3\}$ (for some vertex v) generally has two distinct realizations modulo translations and rotations: supposing that the 3D position of vertices $v-1$, $v-2$, $v-3$ is known, vertex v can be reflected across the unique plane containing the points $v-1$, $v-2$, $v-3$. So if we suppose that the first three atoms have known positions, we can recursively place the remaining atoms by exploring each of the two possible positions at each step (branching step). Those positions which are inconsistent with the distances assigned to the pruning edges are pruned out (pruning step). This yields a method which is exponential in the worst case: if there are no pruning edges, BP yields a binary tree with 2^{n-3} leaf nodes, where n is the number of atoms in the protein. It was recently shown in [16] that this order makes the BP a Fixed-Parameter Tractable (FPT) algorithm.

1.3. Discretization of distance geometry problems

We generalize the backbone order to define an order for K-dimensional spaces: each vertex is adjacent to at least K predecessors [15] (where, specifically for proteins in 3D space, $K = 3$). This number of adjacent predecessors (K) is critical: any fewer, and the solution set might be uncountable in general, since the graph may no longer be rigid [6]; any more, and the corresponding DGP subclass can be solved in polynomial time via trilateration [4].

As discussed above, in protein graphs the adjacent predecessors of any vertex v immediately precede v. This is an important feature: if v has K adjacent vertices that immediately precede v in the order, they are called *contiguous predecessors* of v (those which follow v are called *contiguous successors*). [15]. In summary:

1. the first K vertices in the order form a clique;
2. each vertex with rank greater than K is adjacent to at least K predecessors, exactly K of which are contiguous.

We call the class of DGP instances possessing these orders (and satisfying the strict triangular inequalities on the edge weights [15]) the Discretizable Molecular Distance Geometry Problem in $\mathbb{R}^K$ (^{K}DMDGP), and the orders themselves ^{K}DMDGP *orders*. In [11], the BP algorithm was extended to the ^{K}DMDGP. In [22], it was shown that the BP algorithm could also be used for a larger class of instances, the Discretizable Distance Geometry Problem (DDGP): the DDGP is the subclass of DGP instances for which an order exists (called *DDGP order*) such that Requirement 1 above holds, and a relaxation of Requirement 2 holds, where the K adjacent predecessors need not be contiguous. Both the DDGP and the ^{K}DMDGP are **NP**-hard problems [22,11]. It was shown in [22] that ^{K}DMDGP $\subsetneq$ DDGP (problem P is included in problem Q if the two problems have the same input, and YES (resp. NO) instances of P are also YES (resp. NO) in Q).