Discretization Orders and Distance Geometry

Antonio Mucherino

IRISA, University of Rennes 1, Rennes, France.
antonio.mucherino@irisa.fr

Mots-clés : *distance geometry, discretization order, graph embeddability, de Bruijn graph, combinatorial optimization.*

1 Introduction

The Distance Geometry Problem (DGP) asks whether a simple weighted undirected graph $G = (V, E, d)$ can be embedded in a K-dimensional space [6]. In the applications, K is generally known a priori : one of the most interesting DGP applications is in biology and concerns molecular conformations, where $K = 3$. In such a case, G is defined so that vertices $v \in V$ represent atoms of a molecule, and some possible interatomic distances (weighted edges) are estimated by experimental techniques [7]. The problem of finding a *discretization order*, i.e. an order for the vertices of G that allows for discretizing a DGP instance, is an interesting preprocessing step for the solution of DGPs [4].

In fact, when the DGP can be discretized, its search space can be reduced to a tree, so that the problem becomes combinatorial. A Branch & Prune (BP) algorithm was proposed for the solutions of discretizable DGP instances, which is potentially able to enumerate the entire solution set [5]. For the application in biology, this is of great interest, because more than one molecular conformation may satisfy all available distance constraints, even if only one of them actually represents the conformation that the molecule has.

The definition of a discretization order mainly requires performing three fundamental tasks. First of all, it is necessary that the first K vertices in the discretization order form a clique consisting of exact distances (i.e. of distances not subject to uncertainty). When K is fixed, the problem of finding a K-clique in G can be solved in polynomial time. This initial clique allows for fixing the Cartesian coordinate system where molecular conformations can be identified.

Secondly, for all vertices v having a rank greater than K, the rest of the ordering needs to be constructed. For all such vertices, at least K reference vertices u_1, u_2, ..., u_K, for which $u_i < v$ and $(u_i, v) \in E$, must be available. When only one reference vertex is associated to an uncertain distance, the possible positions in the K-dimensional space for the vertex v belong to two disjoint curves, which either degenerate to two distinct points, or can be successively discretized [3]. This second task can also be performed in polynomial time [4]. First algorithms for the identification of discretization orders only performed these two main tasks.

However, we may not only be interested in finding a discretization order, but rather to select, among all possible orders allowing for the discretization, the ones that optimize some objectives. Such objectives may, for example, allow the BP algorithm to perform better. This last task for the identification of discretization orders is generally the hardest.

2 Consecutivity assumption and de Bruijn graph

The main focus of this work is on orders satisfying the so-called *consecutivity assumption*. The handcrafted order proposed in [5] for the protein backbones is an example of order satisfying this additional assumption. The consecutivity assumption requires that the reference vertices u_1, u_2, ..., u_K, for every $v > K$, are the immediate predecessors of v in the given ordering.

The consecutivity assumption makes it possible to verify in advance the feasibility of all K-cliques composing the order, allowing this way to partially verify in advance the feasibility of the instance. Notice that, while the first two tasks for the identification of discretization orders have polynomial complexity, finding an order for which the consecutivity assumption is satisfied is NP-hard [1].

A discretization order in dimension K satisfying the consecutivity assumption can be seen as a sequence of overlapping $(K+1)$-cliques. $(K+1)$-cliques of G can used for generating a new graph $B = (V_B, A_B)$, where the vertices represent such cliques, and there is an arc from the vertex c to the vertex b of V_A if and only if the two corresponding cliques have K vertices of V in common. When there is an arc (c, b), in fact, there exists an internal ordering for the vertices of c, and an internal ordering for the vertices of b, such that the last K vertices of c coicide with the first K vertices of b. As a consequence, these two cliques can be consecutive in a certain discretization order satisfying the consecutivity assumption.

By definition, graphs B resemble to de Bruijn graphs, generally employed for formulating problems of DNA assembly, where vertices represent sequences of DNA basis, and there is an arc between two sequences if and only if they admit an overlap [2]. The main difference between the standard de Bruijn graph and the graph B is given by the $(K+1)$-cliques, which can actually be seen as sequences of vertices, but with a non-static internal order. In general, every $(K+1)$-clique admits $(K+1)!$ different internal orders.

It is important to remark that additional $(K+1)$-cliques can be generated from K-cliques of G by duplicating one of its vertices. If $C_K = \{v_1, v_2, \ldots, v_K\}$ is a K-clique, then $\{C_K, v_1\}$ is a $(K+1)$-clique containing an edge having weight equal to 0. The use of this kind of cliques is necessary for constructing, for example, the handcrafted order presented in [5]. It allows for generating "bridges" between cliques, when there are no internal orders admitting overlaps with preceding and successive cliques, in any order.

A path on B allows therefore to define a sequence of overlapping $(K+1)$-cliques, i.e. a discretization order for the DGP in dimension K satisfying the consecutivity assumption. In fact, every clique in the path contains the necessary information for computing the set of possible atomic positions for the last atom it contains, in its internal order [3]. Future research will be devoted to the generation of discretization orders for important molecules, such as proteins, by using this novel "pseudo" de Bruijn representation of orders satisfying the consecutivity assumption.

Références

[1] A. Cassioli, O. Günlük, C. Lavor, L. Liberti, *Discretization Vertex Orders in Distance Geometry*, to appear in Discrete Applied Mathematics, 2015.

[2] P.E.C. Compeau, P.A. Pevzner, G. Tesler, *How to Apply de Bruijn Graphs to Genome Assembly*, Nature Biotechnology **29**, 987–991, 2011.

[3] D.S. Gonçalves, A. Mucherino, *Discretization Orders and Efficient Computation of Cartesian Coordinates for Distance Geometry*, Optimization Letters **8**(7), 2111–2125, 2014.

[4] C. Lavor, J. Lee, A. Lee-St.John, L. Liberti, A. Mucherino, M. Sviridenko, *Discretization Orders for Distance Geometry Problems*, Optimization Letters **6**(4), 783–796, 2012.

[5] C. Lavor, L. Liberti, A. Mucherino, *The interval Branch-and-Prune Algorithm for the Discretizable Molecular Distance Geometry Problem with Inexact Distances*, Journal of Global Optimization **56**(3), 855–871, 2013.

[6] L. Liberti, C. Lavor, N. Maculan, A. Mucherino, *Euclidean Distance Geometry and Applications*, SIAM Review **56**(1), 3–69, 2014.

[7] T.E. Malliavin, A. Mucherino, M. Nilges, *Distance Geometry in Structural Biology : New Perspectives.* In : "Distance Geometry : Theory, Methods and Applications", A. Mucherino, C. Lavor, L. Liberti, N. Maculan (Eds.), Springer, 329–350, 2013.