Challenges for extending Discretizable Molecular Distance Geometry to interval data

D. S. Gonçalves,[a] A. Mucherino[b]

[a]*DM, Universidade Federal de Santa Catarina, Florianópolis, Brazil*
douglas@mtm.ufsc.br

[b]*IRISA, University of Rennes 1, Rennes, France*
antonio.mucherino@irisa.fr

Key words: Discretizable Distance Geometry, Embeddability, Interval data

Given a set of distances between pairs of points, the Distance Geometry Problem (DGP) is the one of finding an embedding of the set of points in a K-dimensional Euclidean space. It has several interesting applications such as the Molecular DGP (MDGP) [1], where $K = 3$.

The existence of an embedding in $\mathbb{R}^K$ satisfying a set of *exact* distances can be verified by the Cayley-Menger conditions [2]. In the MDGP, a set of distances is embeddable if and only if all Cayley-Menger determinants of 3 and 4 points have the correct sign(corresponding to the triangular and tetrangular inequalities) and the ones of 5 and 6 points vanish. Another way to verify the embeddability of a set of distances in a generic K-dimensional space is answering whether a partial distance matrix ($D_{ij} = d_{ij}^2$), with *missing* entries, can be completed to a Euclidean distance matrix D. If so, the Cartesian coordinates can be obtained, in polynomial time, by factoring $\mathcal{K}^\dagger(D) = X^T X$, where $\mathcal{K}^\dagger(D)$ is a linear transformation of D and X is a $K \times N$ matrix whose columns are the coordinates of the N points [3].

More recently, a discrete approach was proposed where points are embedded following a predefined *order* which ensures that, for each point, there are at least K distances to previous “non-colinear” reference points [4]. Under the assumptions granted by the order, and in the hypothesis an embedding exists, each point has at most 2 possible positions with respect to the references. Thus, the set of candidate embeddings is *discrete* and the search space becomes a binary tree. Although the worst case complexity of a search in such a tree is exponential, the additional distances, not used in the discretization, can be exploited to prune away infeasible branches and speed up the search. The algorithm that performs this search is called Branch-and-Prune (BP) and its

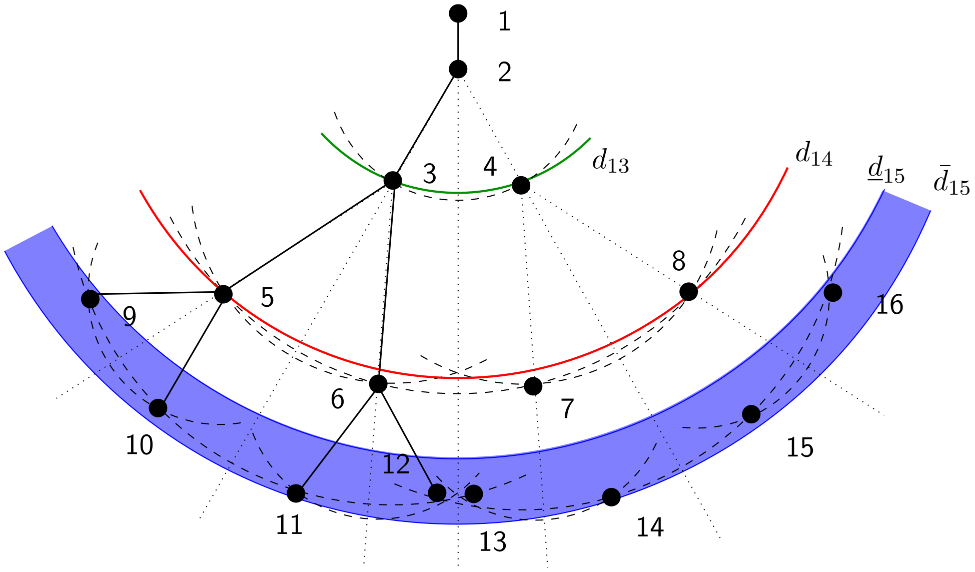

Fig. 1. Illustration of the symmetries in a binary tree for $K = 2$. Although d_{14} prunes out the vertices 6 and 7, the interval distance d_{15} allows any candidate in the fifth layer.

efficiency has been demonstrated in artificial instances related to proteins. Furthermore, by exploiting the symmetries of the search tree [5], it is possible to prove that the tree width is bounded. This explains the polynomial behavior of the BP algorithm in such problems [6].

In practice, however, only the distances provided by covalent geometry can be considered as exact, whereas the information obtained through NMR experiments provides distance bounds. Some preliminary steps for extending the discrete approach to deal with *interval* distances were previously published in [7]. Some interval distances need to be used for the discretization and D equidistant samples are taken from each of them in order to generate $2 \times D$ candidate positions. As a consequence, the search tree is no longer binary: the extension of BP to deal with interval distances is named *interval* BP (iBP).

Recent computational experiments [8, 9] showed that the difficulties encountered by iBP are due to the following main reasons:

- the sampled distances are taken independently in each layer of the tree and, in particular for D small, it is *not* likely that they satisfy the embeddability conditions;
- interval pruning distances affect the boundness of the tree width. Even if the tree is binary, an interval pruning distance may allow much more than two feasible nodes per layer (see Figure 1), possibly leading to a combinatorial explosion.

The latter difficulty can be mitigated by introducing additional pruning criteria, for instance, based on chemical-physical properties of the molecule [10].

In this work, we handle the first issue. By taking into account the regular structure of the protein backbone, we propose heuristics for *joint sampling* from interval distances between hydrogens atoms. As a preprocessing, we apply bound smoothing [2] and Cayley-Menger inequalities to reduce the interval distances. Then, in the first phase, we obtain samples for the distances related to hydrogens by metrization [2] or solving a matrix completion problem [3] involving hydrogens only. In the second phase, the obtained distances between hydrogens feedback the iBP algorithm to guide the sample selection. In each layer, we shrink the discretization interval by intersecting the previous pruning distances [9]. This process provides a set of feasible intervals from where the samples are taken. We select as sample distances, in the reduced discretization intervals, the closest ones to the candidate distance obtained by the first phase. Therefore, each sampled discretization distance is compatible with the (so far) available pruning distances and minimize the deviation to the distance suggested by phase one.

The presented approach represents a little step towards the solution of Discretizable MDGPs with interval data where the sampling phase of iBP is critical, and a joint sampling is necessary due to the dependency between interval discretization distances, imposed by the embeddability conditions.

References

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

[2] T. F. Havel, I. D. Kuntz, G. M. Crippen. *The Theory and Practice of Distance Geometry*, Bulletin of Mathematical Biology **45**(5), 665–720, 1983.

[3] B. Alipanahi, N. Krislock, A. Ghodsi, H. Wolkowicz, L. Donaldson, and M. Li. *Determining protein structures from NOESY distance constraints by semidefinite programming.* Journal of Computational Biology, **20**(4), 296–310, 2013.

[4] C. Lavor, L. Liberti, N. Maculan, A. Mucherino. *The discretizable molecular distance geometry problem*, Computational Optimization and Applications **52**, 115–146, 2012.

[5] L. Liberti, B. Masson, J. Lee, C. Lavor, A. Mucherino. *On the number of realizations of certain henneberg graphs arising in protein conformation* Discrete Applied Mathematics, **165**, 213–232, 2014.

[6] L. Liberti, C. Lavor, A. Mucherino. *The discretizable molecular distance geometry problem seems easier on proteins*, In A. Mucherino, C. Lavor, L. Liberti, and N. Maculan, editors, *Distance Geometry*, pages 47–60. Springer New York, 2013.

[7] 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.

[8] D. S. Gonçalves A. Mucherino. *Discretization orders and efficient computation of Cartesian coordinates for distance geometry*, to appear in Optimization Letters, 2014.

[9] D. S. Gonçalves, A. Mucherino, C. Lavor. *An Adaptive Branching Scheme for the Branch and Prune Algorithm applied to Distance Geometry* , to appear in IEEE Conference Proceedings, Federated Conference on Computer Science and Information Systems (FedCSIS14), Workshop on Computational Optimization (WCO14), Warsaw, Poland, September, 2014.

[10] D. S. Gonçalves, A. Mucherino, C. Lavor. *Energy-based pruning devices for the BP algorithm applied to Distance Geometry* , in IEEE Conference Proceedings, Federated Conference on Computer Science and Information Systems (FedCSIS13), Workshop on Computational Optimization (WCO13), pages 341–346. Krakow, Poland, September, 2013.