On the polynomiality of finding ^{K}DMDGP re-orders

Carlile Lavor[a,1], Michael Souza[b,2], Luiz Mariano Carvalho[c], Leo Liberti[d,*]

[a] *University of Campinas (IMECC-UNICAMP), 13081-970, Campinas - SP, Brazil*
[b] *Federal University of Ceará (DEMA-UFC), 60020-181, Fortaleza - CE, Brazil*
[c] *Rio de Janeiro State University (IME-UERJ), 20559-900, Rio de Janeiro - RJ, Brazil*
[d] *CNRS LIX, École Polytechnique, 91128 Palaiseau, France*

Abstract

In [6], the complexity of finding ^{K}DMDGP re-orders was stated to be **NP**-complete by inclusion, which fails to provide a complete picture. In this paper we show that this problem is indeed **NP**-complete for $K = 1$, but it is in **P** for each fixed $K \geq 2$.

Keywords: Distance Geometry, Discretizable Molecular Distance Geometry Problem, Vertex Orders

1. Introduction

The fundamental problem in Distance Geometry (DG) is the DG Problem (DGP): given an integer $K > 0$ and a simple, undirected, non-negatively edge-weighted graph $G = (V, E, d)$, with $d : E \to \mathbb{R}_+$, find positions in $\mathbb{R}^K$ for each vertex such that each edge, drawn as a segment, has length equal to the weight [25, 26, 28]. The set of positions of all the vertices in V is called a *realization* of G. Many variants replace equality with inequalities to address data measurement error and noise [1, 2, 7, 10, 14, 18, 21, 33, 34]. The DGP has applications to many fields of science and engineering, including clock synchronization protocols, sensor network localization, robotics, nanostructures, and protein structure determination [3, 4, 9, 11, 19, 31].

Most of the solution methods for the DGP on arbitrary graphs consist of search algorithms in continuous space [23], but if an appropriate vertex order is given, the DGP solution space becomes discrete [13, 15, 30].

Our motivation is based on a particular subclass of DGP graphs which arise in the modelling of proteins, where $K = 3$ [8, 32]. More specifically, we look at graphs which are models of protein backbones. The edge set of these graphs, related to the pairs of atoms whose distances are known theoretically and experimentally, contains:

*Corresponding author, partly sponsored by the EU H2020 Programme ETN "MINOA" grant agreement n. 764759.

Email addresses: clavor@ime.unicamp.br (Carlile Lavor), souza.michael@gmail.com (Michael Souza), luizmc@gmail.com (Luiz Mariano Carvalho), liberti@lix.polytechnique.fr (Leo Liberti)

[1]Partly supported by FAPESP, CNPq

[2]Partly supported by FAPESP

1. a Hamiltonian path which is used to number the vertices from v_1 to v_n, where $n = |V|$;
2. a clique defined on the first three vertices v_1, v_2, v_3 (in general, up to v_K);
3. for each vertex v_i in the Hamiltonian path, $i > 3$, edges connecting v_i to its three immediate predecessors, namely vertices $v_{i-1}, v_{i-2}, v_{i-3}$ (in general, for each $i > K$, edges connecting v_i to its K immediate predecessors);
4. possibly other edges not described by (1)-(3).

Assuming that the strict triangular inequality is satisfied for $v_{i-3}, v_{i-2}, v_{i-1}$, $i > 3$, these graphs are rigid in $\mathbb{R}^3$ (a similar generalized statement holds for K and strict simplex inequalities). The associated DGP subclass is called the *Discretizable Molecular Distance Geometry Problem* (DMDGP) [16, 17]. In [27], it is proved that the number of its realizations is almost always a power of two, implying interesting symmetric properties in the solution space [12, 24, 29], where a Branch-and-Prune algorithm can be applied to find all these realizations [5, 22].

We now look at an auxiliary problem: given an arbitrary graph, does its edge set satisfy the conditions (1)-(4) given above? In other words, can we find a vertex order such that (1)-(3) hold? The problem, informally known as "*DMDGP order*", was named *Contiguous Trilateration Ordering Problem* (CTOP) in [6], where it was shown (by reduction from *Hamiltonian Path*) that CTOP is **NP**-complete.

In order to address this computational complexity issue arising in protein structure determination applications of the DGP, "hand-crafted" vertex orders were proposed based on the atomic sequences in protein backbones [18, 20]. The novelty introduced by such orders is that they allow the repetition of vertices in the order. More precisely, the vertex of rank i in the order (for $i > 3$) also has the possibility of being the same vertex having rank less than $i - 2$. Such orders are called *repetition orders* (or *re-orders*).

We generalize the fact that proteins exist in at most three spatial dimensions by replacing the number three with a general integer $K > 0$. In this setting, the problem of finding a re-order relative to K in an arbitrary graph is called the *Re-Order Problem* (ReOP). In [6, §3.3] it was stated that "because a re-order which never repeats any vertices is a ^{K}DMDGP order, the ReOP is also **NP**-complete for any fixed K by restriction to the CTOP". This sentence is wrong: the correct sentence is "the ReOP is **NP**-complete by inclusion of the case when $K = 1$". In finding a re-order, since we are free to repeat vertex v_i at rank $i + K$, it follows that re-orders do not need to be Hamiltonian paths. The ReOP is **NP**-complete for $K = 1$ because repeating vertex v_i at rank $i + 1$ is equivalent to "failing to progress in the construction of the order". Therefore, for $K = 1$, the reduction from *Hamiltonian Path* to ReOP is trivial. This implies that ReOP is **NP**-complete for $K = 1$, and that it is **NP**-complete by inclusion of this case. For higher values of K, however, it turns out that re-orders can be found in polynomial time, including the case of interest for proteins $K = 3$.

2. Finding a ^{K}DMDGP re-order in polynomial time for $K > 1$

We start by giving the precise definition of a ^{K}DMDGP re-order for a graph G and the definition of an auxiliary K-clique incidence graph G_K derived from G. We consider $K > 1$.

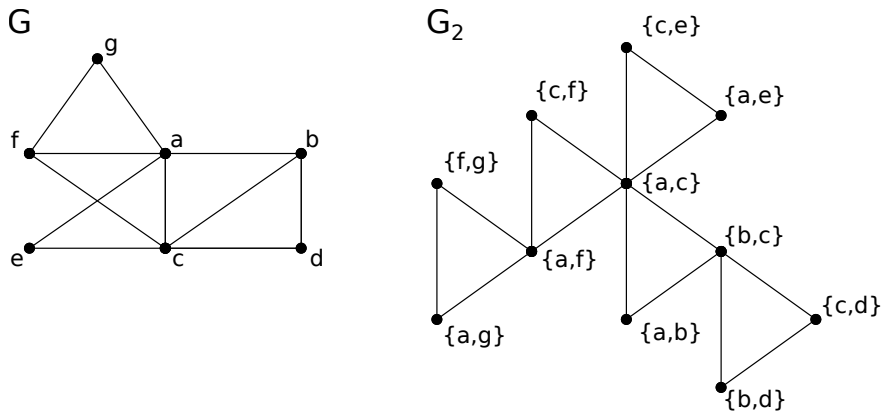

Figure 1: An example of G and G_2.

Definition 1. Given a simple undirected graph $G = (V, E)$, a ^{K}DMDGP re-order for G is a surjective function $r : \{1, \ldots, m\} \to V$, with length $m \in \mathbb{N}$ bounded by a polynomial in $|V|$, such that (we simplify the notation using r_i instead of $r(i)$):

1. $\{r_1, \ldots, r_K\}$ defines a K-clique in G;
2. For $i \in \{K+1, \ldots, m\}$, $\{r_{i-K+1}, r_i\}, \{r_{i-K+2}, r_i\}, \ldots, \{r_{i-1}, r_i\} \in E$;
3. For $i \in \{K+1, \ldots, m\}$, $\{r_{i-K}, r_i\} \in E$ or $r_{i-K} = r_i$.

Definition 2. Given a simple undirected graph $G = (V, E)$, the graph $G_K = (V_K, E_K)$ is defined as follows:

1. $\mathfrak{u} \in V_K$ iff $\mathfrak{u}$ is a set defined by the vertices of a K-clique in G;
2. $\{\mathfrak{u}, \mathfrak{v}\} \in E_K$ iff $\mathfrak{u} \cup \mathfrak{v}$ is a set given by the vertices of a $(K+1)$-clique in G.

In Figure 1, we give an example of G and G_K, for $K = 2$. The next theorem identifies ^{K}DMDGP re-orders for G with particular walks in G_K.

Theorem 1. *Given a simple undirected graph $G = (V, E)$, there is a ^{K}DMDGP re-order r in G if and only if there exists a walk $\gamma = (\gamma_1, \ldots, \gamma_{m-K+1})$ in $G_K = (V_K, E_K)$ such that $\gamma_i \in V_K$ for all $i \leq m - K + 1$ and $\bigcup_{i \leq m-K+1} \gamma_i = V$.*

Proof. Given a ^{K}DMDGP re-order r in G, with length $m \in \mathbb{N}$, we define γ_i, for $i \in \{1, \ldots, m-K+1\}$, in the following way:

$$\begin{aligned} \gamma_1 &= \{r_1, r_2, \ldots, r_K\}, \\ \gamma_2 &= \{r_2, r_3, \ldots, r_{K+1}\}, \\ &\vdots \\ \gamma_{m-K+1} &= \{r_{m-K+1}, r_{m-K+2}, \ldots, r_m\}. \end{aligned}$$

Since $\gamma_i = \{r_i, r_{i+1}, \ldots, r_{i+K-1}\}$ defines a K-clique in G (by definition of ^{K}DMDGP re-order), it follows that $\gamma_i \in V_K$ for each $i \in \{1, \ldots, m-K+1\}$. Moreover, since $\gamma_i \neq \gamma_{i+1}$, by definition of G_K we have $\{\gamma_i, \gamma_{i+1}\} \in E_K$, which implies that γ is a walk in G_K such that $\bigcup_i \gamma_i = V$ (because r is a surjection).

Conversely, let us suppose that there exists a walk γ of length p in G_K with $\bigcup\limits_{i \leq p} \gamma_i = V$. Consider an arbitrary ordering on γ_1 given by $\gamma_1 = (\gamma_1^1, \gamma_1^2, \ldots, \gamma_1^K)$. For $i \in \{2, \ldots, p\}$ and $j \in \{1, \ldots K\}$ define:

$$\gamma_i^j = \begin{cases} \gamma_{i-1}^j & \text{if } \gamma_{i-1}^j \in \gamma_i \\ \gamma_i^* & \text{otherwise,} \end{cases}$$

where γ_i^* is the unique element of $\gamma_i \smallsetminus \gamma_{i-1}$. We claim that the function $r : \{1, \ldots, pK\} \to V$ defined by

$$r_i = \gamma_{\lceil i/K \rceil}^{i-(\lceil i/K \rceil - 1)K}$$

is a ^{K}DMDGP re-order. We will show it satisfies the requirements of the definition:

1. $\{r_1, \ldots, r_K\}$ is a K-clique in G: by definition $\{r_1, \ldots, r_K\} = \{\gamma_1^1, \gamma_1^2, \ldots, \gamma_1^K\} = \gamma_1$, where γ_1 is a K-clique in G by definition of G_K.
2. For $\ell \in \{K+1, \ldots, pK\}$, $\{r_{\ell-K+1}, r_\ell\}, \{r_{\ell-K+2}, r_\ell\}, \ldots, \{r_{\ell-1}, r_\ell\} \in E$: by definition of r, $\{r_{\ell-K+1}, r_{\ell-K+2}, \ldots, r_\ell\} \subset \gamma_{\lceil \ell/K \rceil - 1} \cup \gamma_{\lceil \ell/K \rceil}$, for $\ell \in \{K+1, \ldots, pK\}$, which implies that $\{r_{\ell-j}, r_\ell\} \in E$ for $j \in \{1, \ldots, K-1\}$ by definition of G_K.
3. For $\ell \in \{K+1, \ldots, pK\}$, $\{r_{\ell-K}, r_\ell\} \in E$ or $r_{\ell-K} = r_\ell$: again by definition of r and G_K, for $\ell = K+1, \ldots, pK$ we have that $r_{\ell-K} = r_\ell$ if $r_\ell \in \gamma_{\lceil \ell/K \rceil - 1} \cap \gamma_{\lceil \ell/K \rceil}$, and $\{r_{\ell-K}, r_\ell\} \in E$ if $r_{\ell-K} \in \gamma_{\lceil \ell/K \rceil - 1}$ and $r_\ell \in \gamma_{\lceil \ell/K \rceil}$.

This concludes the proof. □

The next result states the relationship between ^{K}DMDGP re-orders in G and the connectivity of G_K.

Corollary 1. *There is a ^{K}DMDGP re-order in a simple undirected graph $G = (V, E)$ if and only if there exists a connected component of $G_K = (V_K, E_K)$ whose vertex union is V.*

Proof. This follows immediately from the Theorem 1. □

The last two results ensure that ^{K}DMDGP re-orders can be found in polynomial time.

Lemma 1. *Given a simple undirected graph $G = (V, E)$, the graph $G_K = (V_K, E_K)$ can be constructed in $O(K^3|V|^{K+1})$ steps.*

Proof. Constructing G_K requires constructing its edge set E_K. Each vertex of each $(K+1)$-clique in G yields an incidence of two K-cliques in G. By definition, this means that there are $K+1$ edges in E_K for each $(K+1)$-clique in G. In the worst case, the number of $(K+1)$-cliques in G is $\binom{|V|}{K+1} = O(|V|^{K+1})$, which yields an $O(K|V|^{K+1})$ worst-case time complexity bound. Furthermore, considering the time for verifying whether a set of vertices is a clique is $O(K^2)$, we conclude that the worst case time complexity for constructing G_K is $O(K^3|V|^{K+1})$. □

Theorem 2. *A ^{K}DMDGP re-order for $G = (V, E)$ can be found in $O(|V|^{2K})$ steps.*

Proof. Using DFS or BFS, we find connected components in $O(n + m)$ steps for a graph with n vertices and m edges; the worst case is on dense graphs where m is $O(n^2)$. G_K has at worst $\binom{|V|}{K}$ vertices and $O(\binom{|V|}{K}^2)$ edges. As mentioned above, $\binom{|V|}{K}$ is dominated by $O(|V|^K)$, so this step of the algorithm yields a $O(|V|^{2K})$ time bound. Finding a spanning tree requires $O(m+n\log n)$ steps using Prim's algorithm with Fibonacci heaps; in the worst case the $O(m) = O(n^2)$ term dominates, so that also yields a $O(|V|^{2K})$ time bound. A walk is obtained from a tree by replacing each edge with two antiparallel arcs, and then exploring the corresponding Eulerian cycle, which can be done in $O(n) = O(|V|^K)$ steps. Constructing G_K from G requires $O(K^3|V|^{K+1})$ steps by Lemma 1. The whole algorithm has a worst-case time complexity $O(K^3|V|^{K+1} + 2|V|^{2K} + |V|^K) = O(|V|^{2K})$, since $K < |V|$. □

In conclusion, we have the following theorem which summarizes our discussion.

Theorem 3. *Given a simple undirected graph $G = (V, E)$, the problem of finding a ^{K}DMDGP re-order is* **NP***-complete for $K = 1$, and in* **P** *for each fixed $K \geq 2$.*

Acknowledgements

We thank the Brazilian research agencies CNPq and FAPESP for support. We thank A. Mucherino, D. Gonçalves, and R. Marques for useful discussions, as well as two anonymous referees for valuable suggestions.

References

[1] R. Alves and C. Lavor, Geometric algebra to model uncertainties in the discretizable molecular distance geometry problem, Advances in Applied Clifford Algebra, 27 (2017), 439-452.

[2] R. Alves, C. Lavor, C. Souza, and M. Souza, Clifford algebra and discretizable distance geometry, Mathematical Methods in the Applied Sciences, 41 (2018), 3999–4346.

[3] S. Billinge, P. Duxbury, D. Gonçalves, C. Lavor, and A. Mucherino, Assigned and unassigned distance geometry: applications to biological molecules and nanostructures, 4OR, 14 (2016), 337–376.

[4] S. Billinge, P. Duxbury, D. Gonçalves, C. Lavor, and A. Mucherino, Recent results on assigned and unassigned distance geometry with applications to protein molecules and nanostructures, Annals of Operations Research, 271 (2018), 161–203.

[5] R. Carvalho, C. Lavor, and F. Protti, Extending the geometric build-up algorithm for the molecular distance geometry problem, Information Processing Letters, 108 (2008), 234–237.

[6] A. Cassioli, O. Gunluk, C. Lavor, and L. Liberti, Discretization vertex orders in distance geometry, Discrete Applied Mathematics, 197 (2015), 27–41.

[7] A. Cassioli, B. Bordiaux, G. Bouvier, A. Mucherino, R. Alves, L. Liberti, M. Nilges, C. Lavor, and T. Malliavin, An algorithm to enumerate all possible protein conformations verifying a set of distance constraints, BMC Bioinformatics, 16 (2015), 16–23.

[8] V. Costa, A. Mucherino, C. Lavor, A. Cassioli, L. Carvalho, and N. Maculan, Discretization orders for protein side chains, Journal of Global Optimization, 60 (2014), 333-349.

[9] G. Crippen and T. Havel, Distance Geometry and Molecular Conformation, Wiley, New York, (1988).

[10] C. Dambrosio, V. Ky, C. Lavor, L. Liberti, and N. Maculan, New error measures and methods for realizing protein graphs from distance data. Discrete & Computational Geometry, 57 (2017), 371–418.

[11] B. Donald, Algorithms in Structural Molecular Biology, MIT Press, Boston, (2011).

[12] F. Fidalgo, D. Gonçalves, C. Lavor, L. Liberti, and A. Mucherino, A symmetry-based splitting strategy for discretizable distance geometry problems, Journal of Global Optimization, 71 (2018), 717–733.
[13] D. Gonçalves and A. Mucherino, Discretization orders and efficient computation of Cartesian coordinates for distance geometry, Optimization Letters 8 (2014), 2111-2125.
[14] D. Gonçalves, A. Mucherino, C. Lavor, and L. Liberti, Recent advances on the interval distance geometry problem, Journal of Global Optimization, 69 (2017), 525-545.
[15] C. Lavor, J. Lee, A. Lee-St. John, L. Liberti, A. Mucherino, and M. Sviridenko, Discretization orders for distance geometry problems, Optimization Letters, 6 (2012), 783–796.
[16] C. Lavor, L. Liberti, N. Maculan, and A. Mucherino, The discretizable molecular distance geometry problem, Computational Optimization and Applications, 52 (2012), 115–146.
[17] C. Lavor, L. Liberti, N. Maculan, and A. Mucherino, Recent advances on the discretizable molecular distance geometry problem, European Journal of Operational Research, 219 (2012), 698-706.
[18] C. Lavor, L. Liberti, and A. Mucherino, The interval branch-and-prune algorithm for the discretizable molecular distance geometry problem with inexact distances, Journal of Global Optimization, 56 (2013), 855–871.
[19] C. Lavor, L. Liberti, W. Lodwick, and T. Mendonça da Costa, An Introduction to Distance Geometry applied to Molecular Geometry, SpringerBriefs, Springer, New York, (2017).
[20] C. Lavor, L. Liberti, B. Donald, B. Worley, B. Bardiaux, T. Malliavin, and M. Nilges, Minimal NMR distance information for rigidity of protein graphs, Discrete Applied Mathematics, 256 (2019), 91–104.
[21] C. Lavor and R. Alves, Oriented conformal geometric algebra and the molecular distance geometry problem, Advances in Applied Clifford Algebra, 29 (2019), 1-19.
[22] L. Liberti, C. Lavor, and N. Maculan, A branch-and-prune algorithm for the molecular distance geometry problem, International Transactions in Operational Research, 15 (2008), 1–17.
[23] L. Liberti, C. Lavor, A. Mucherino, and N. Maculan, Molecular distance geometry methods: from continuous to discrete, International Transactions in Operational Research, 18 (2010), 33–51.
[24] L. Liberti, C. Lavor, J. Alencar, and G. Resende, Counting the number of solutions of ^{K}DMDGP instances, Lecture Notes in Computer Science, 8085 (2013), 224–230.
[25] L. Liberti, C. Lavor, N. Maculan, and A. Mucherino, Euclidean distance geometry and applications, SIAM Review, 56 (2014), 3–69.
[26] L. Liberti and C. Lavor, Six mathematical gems from the history of distance geometry, International Transactions in Operational Research, 23 (2016), 897–920.
[27] L. Liberti, B. Masson, J. Lee, C. Lavor, and A. Mucherino, On the number of realizations of certain Henneberg graphs arising in protein conformation, Discrete Applied Mathematics, 165 (2014), 213–232.
[28] L. Liberti and C. Lavor, Euclidean Distance Geometry: An Introduction, Springer, New York, (2017).
[29] A. Mucherino, C. Lavor, and L. Liberti, Exploiting symmetry properties of the discretizable molecular distance geometry problem, Journal of Bioinformatics and Computational Biology, 10 (2012), 1242009(1-15).
[30] A. Mucherino, C. Lavor, and L. Liberti, The discretizable distance geometry problem, Optimization Letters, 6 (2012), 1671–1686.
[31] A. Mucherino, C. Lavor, L. Liberti, and N. Maculan, eds., Distance Geometry: Theory, Methods, and Applications, Springer, New York, (2013).
[32] S. Sallaume, S. Martins, L. Ochi, W. Gramacho, C. Lavor, and L. Liberti, A discrete search algorithm for finding the structure of protein backbones and side chains, International Journal of Bioinformatics Research and Applications, 9 (2013), 261–270.
[33] M. Souza, A. Xavier, C. Lavor, and N. Maculan, Hyperbolic smoothing and penalty techniques applied to molecular structure determination, Operations Research Letters, 39 (2011), 461-465.
[34] M. Souza, C. Lavor, A. Muritiba, and N. Maculan, Solving the molecular distance geometry problem with inaccurate distance data, BMC Bioinformatics, 14 (2013), S71–S76.