Amazing results in the mathematics of proteins

Leo Liberti

LIX, École Polytechnique, F-91128 Palaiseau, France
Email:liberti@lix.polytechnique.fr

May 7, 2011

Abstract

This report describes the two-hour mini-course given by myself at the Pretty Structures 2011 workshop (http://www.lix.polytechnique.fr/~liberti/pretty_structures). This work is due to a research team also including Carlile Lavor, Jon Lee, Benoıt Masson and Antonio Mucherino. The proofs and mistakes in this presentations are, however, entirely mine.

The problem I treat here is the Distance Geometry Problem (DGP): we are given a simple undirected graph $G = (V, E)$ with a real-valued edge weight function $d : E \to \mathbb{R}_{\geq 0}$ and we look for an embedding $x : V \to \mathbb{R}^K$ of the vertices of G which preserves the Euclidean distances assigned to each edge, i.e. $\forall \{u, v\} \in E \; \|x_u - x_v\| = d_{uv}$. I shall limit the problem to consist of a set of instances describing protein graphs.

Past computational experience showed us that protein graphs always had a set of solutions of cardinality either empty or a power of two; but we could manually construct counterexamples to this conjecture.

I find the results contained here amazing because they show that essentially the conjecture holds *notwithstanding* the counterexamples! Furthermore, an amazing consequence (not discussed here) of this "power of two" property is that in we can find protein conformations in polynomial time.

1 Introduction

Pharmaceutical companies manufacture drugs, and in order to do so they must know the biological function of several proteins. It turns out that the function of proteins is strongly linked to the shape it has in 3D space [13]. In order to determine this shape we dispose of a certain amount of information. In this work I assume some general knowledge on the structure of each protein, some chemical information about certain inter-atomic distances, and experimental information about another set of inter-atomic distances. More precisely, I shall make the following assumptions:

- each protein can be decomposed into a sequence of atoms with a linear order, called the *backbone*, and several sets (called *side chains*) of variously interconnected atoms which are connected to the backbone via only one link between each side chain and the backbone;
- we know the distances between each atom on the backbone and its two preceding (and hence two subsequent) atoms in the order;
- we can measure the distances between each atom i on the backbone and the atom ranked $i - 3$ in the order using Nuclear Magnetic Resonance (NMR).

Because of the first assumption, the problem of finding protein embeddings can be decomposed into finding embeddings of the backbone and, separately, of each side chain. In this work I only treat the former. The second and third assumptions will allow me to describe the class of graphs for which I mean to find embeddings.

2 The problem

I shall consider the following decision problem [6]:

> K-Discretizable Molecular Distance Geometry Problem ($^{\mathsf{K}}$DMDGP). Given a positive integer K and a simple weighted undirected graph $G = (V, E, d)$ where $d : E \to \mathbb{R}_+$, V is ordered so that $V = [n] = \{1, \ldots, n\}$ and the following assumptions hold:
>
> 1. for all $v > K$ and $u \in V$ with $1 \leq v - u \leq K$, $\{u, v\} \in E$ (Discretization)
> 2. for all $v > K$, E contains all edges $\{u, w\}$ with $u \neq w \in U_v = \{u \in V \mid 1 \leq v - u \leq K\}$, and the distances d_{uw} with $u \neq w \in U_v$ obey the strict simplex inequalities [1] (Strict Simplex Inequalities),
>
> and given an embedding $x' : [K] \to \mathbb{R}^K$, is there an embedding $x : V \to \mathbb{R}^K$ extending x', such that
>
> $$\forall \{u, v\} \in E \quad \|x_u - x_v\| = d_{uv} \; ? \tag{1}$$

Note that the strict simplex inequalities in $\mathbb{R}^3$ reduce to the strict triangular inequalities $d_{v-3,v-1} < d_{v-3,v-2} + d_{v-2,v-1}$. An embedding x extends an embedding x' if x' is a restriction of x; an embedding is feasible if it satisfies (1). Other related problems also exist in the literature, such as the Discretizable Distance Geometry Problem (DDGP) [11], where the Discretization axiom is relaxed to require that each vertex $v > K$ has at least K adjacent predecessors. The results in these notes, however, only refer to the $^{\mathsf{K}}$DMDGP.

This problem models the protein conformation problem described in the introduction. For any atom $v \in V$, the distances $d_{v-1,v}$ and $d_{v-2,v-1}$ are known because they refer to covalent bonds. Furthermore, the angle between $v-2$, $v-1$ and v is known because it is adjacent to two covalent bonds, which implies that $d_{v-2,v}$ is also known by triangular geometry. In general, the distance $d_{v-3,v}$ is smaller than 5Å and can therefore be assumed to be known by NMR experiments; in practice, there are ways to find atomic orders which ensure that $d_{v-3,v}$ is known [7]. There is currently no known protein with $d_{v-3,v-1}$ being *exactly equal* to $d_{v-3,v-2} + d_{v-2,v-1}$ [8].

2.1 Probability 1 statements

Statement such as "$\forall p \in P \; F(p)$ holds with probability 1", for some uncountable set P and valid sentence F, actually mean that there is a Lebesgue-measurable $Q \subseteq P$ with Lebesgue measure 1 w.r.t. P such that $\forall p \in Q \; F(p)$ holds. This notion is less restrictive than genericity based on algebraic independence [2].

2.2 The DMDGP axioms

I shall give explanations and examples for the cases $K = 2$ and $K = 3$ which are easier to visualize. The Discretization axiom guarantees that the locus of the points embedding v in $\mathbb{R}^3$ is the intersection of the three spheres centered at $v-3, v-2, v-1$ with radii $d_{v-3,v}, d_{v-2,v}, d_{v-1,v}$. If this intersection is non-empty, then it contains two points (Fig. 1, left) apart from a set of Lebesgue measure 0 where it may contain either one point or infinitely many (Fig. 1, right). The role of the Strict Simplex Inequalities axiom is to prevent the latter case of infinitely many points. As such, one might actually dispense with this axiom altogether and simply state that all the results hold with probability 1. Remark that if the intersection of the three spheres is empty, then the instance is a NO one.

The Discretization axiom allows the solution of $^{\mathsf{K}}$DMDGP instances using a recursive algorithm called Branch-and-Prune (BP) [8]: at level v, the search is branched according to the (at most two)

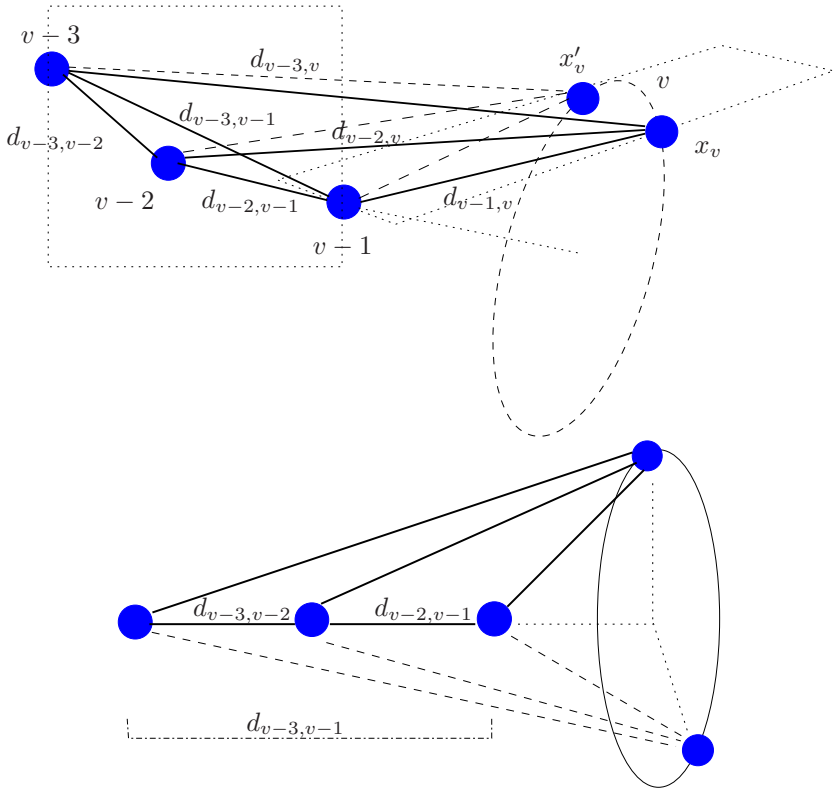

Figure 1: Locus of the intersection of three spheres: exactly two points (above) with $d_{v-3,v-1} < d_{v-3,v-2} + d_{v-2,v-1}$ and uncountably many (below) with $d_{v-3,v-1} = d_{v-3,v-2} + d_{v-2,v-1}$.

possible positions for v. The BP generates a (partial) binary search tree of height n, each full branch of which represents a feasible embedding for the given graph.

3 The BP algorithm

For all $v \in V$ we let $N(v) = \{u \in V \mid \{u, v\} \in E\}$ be the set of vertices *adjacent* to v. An embedding of a subgraph of G is called a *partial embedding* of G. We denote by X the set of embeddings (modulo congruences) solving a $^{\mathsf{K}}$DMDGP instance.

The BP algorithm exploits the edges guaranteed by the DISCRETIZATION axiom in order to search a discrete set: vertex v can be placed in at most two possible positions (the intersection of K spheres in $\mathbb{R}^K$). Each is tested in turn and the procedure called recursively for each feasible position. The BP exploits all other edges in the graph in order to prune some branches: a position might be feasible with respect to the distances to the K immediate predecessors $v-1, \ldots, v-K$, but not necessarily with distances to other adjacent predecessors.

For a partial embedding $\bar{x}$ of G and $\{u, v\} \in E$ let $S^{\bar{x}}_{uv}$ be the sphere centered at x_u with radius d_{uv}. The BP algorithm is BP$(K+1, x', \emptyset)$ (see Alg. 1), where x' is the initial embedding of the first K vertices mentioned in the problem definition. By the $^{\mathsf{K}}$DMDGP axioms, $|T| \leq 2$. At termination, X contains all embeddings (modulo congruences) extending x' [8, 6].

Algorithm 1 BP(v, $\bar{x}$, X)

Require: A vtx. $v \in V \setminus [K]$, a partial embedding $\bar{x} = (x_1, \ldots, x_{v-1})$, a set X.

1: $T = \bigcap_{\substack{u \in N(v) \\ u < v}} S^{\bar{x}}_{uv}$;

2: $\forall p \in T$ { **let** $x = (\bar{x}, p)$; **if** $(v = n)$ $X \leftarrow X \cup \{x\}$ **else** BP($v + 1$, x, X) }.

3.1 Chirality

Embeddings $x \in X$ can be represented by sequences $\chi(x) \in \{-1, 1\}^n$ with: (i) $\chi(x)_i = 1$ for all $i \leq K$; (ii) for all $i > K$, $\chi(x)_i = -1$ if $ax_i < a_0$ and $\chi(x)_i = 1$ if $ax_i \geq a_0$, where $ax = a_0$ is the equation of the hyperplane through $x_{i-K}, \ldots, x_{i-1}$. For an embedding $x \in X$, $\chi(x)$ is the *chirality* [3, 12] of x (the formal definition of chirality actually states $\chi(x)_0 = 0$ if $ax_i = a_0$, but since this case holds with probability 0, we do not consider it here).

3.2 Advertisement

The BP (Alg. 1) can be run to termination to find all possible embeddings of G, or stopped after the first leaf node at level n is reached, in order to find just one embedding of G. In the last few years we have conceived and described several BP variants targeting different problems [5], including, very recently, problems with interval-type uncertainties on some of the distance values [7]. The BP algorithm is currently the only method which is able to find all incongruent embeddings for a given protein backbone. Compared to continuous search algorithms (e.g. [10]), the performance of the BP algorithm is impressive from the point of view of both efficiency and reliability.

You can download an open-source BP implementation from `http://www.antoniomucherino.it/download/mdjeep/mdjeep-0.1.tar.gz`.

4 BP search trees with bounded width

We partition E into the sets $E_D = \{\{u, v\} \mid |v - u| \leq K\}$ and $E_P = E \setminus E_D$. We call E_D the *discretization edges* and E_P the *pruning edges*. Discretization edges guarantee that a DGP instance is in the $^{\mathsf{K}}$DMDGP. Pruning edges are used to reduce the BP search space by pruning its tree. In practice, pruning edges might make the set T in Alg. 1 have cardinality 0 or 1 instead of 2. We assume G is a YES instance of the $^{\mathsf{K}}$DMDGP.

4.1 The discretization group

Let $G_D = (V, E_D, d)$ and X_D be the set of embeddings of G_D; since G_D has no pruning edges, the BP search tree for G_D is a full binary tree and $|X_D| = 2^{n-K}$. The discretization edges arrange the embeddings so that, at level ℓ, there are $2^{\ell-K}$ possible positions for the vertex v with rank ℓ. We assume that $|T| = 2$ (see Alg. 1) at each level v of the BP tree, an event which, in absence of pruning edges, happens with probability 1 — thus many results in this section are stated with probability 1. Let therefore $T = \{x_v, x'_v\}$ be the two possible embeddings of v at a certain recursive call of Alg. 1 at level v of the BP tree; then because T is an intersection of K spheres, x'_v is the reflection of x_v through the hyperplane defined by $x_{v-K}, \ldots, x_{v-1}$. Denote this reflection operator by R^v_x.

4.1 Theorem (Cor. 4.5 and Thm. 4.8 in [9])
With probability 1, for all $v > K$ and $u < v - K$ there is a set H^{uv}, with $|H^{uv}| = 2^{v-u-K}$, of real positive values such that for each $x \in X$ we have $\|x_v - x_u\| \in H^{uv}$. Furthermore, $\forall x \in X$ $\|x_v - x_u\| =$

$\|R_x^{u+K}(x_v) - x_u\|$ *and* $\forall x' \in X$, *if* $x'_v \notin \{x_v, R_x^{u+K}(x_v)\}$ *then* $\|x_v - x_u\| \neq \|x'_v - x_u\|$.

Proof. Sketched in Fig. 2 for $K = 2$; the solid circles at levels $3, 4, 5$ mark equidistant levels from 1. The dashed circles represent the spheres S_{uv}^x (see Alg. 1). Intuitively, two branches from level 1 to level 4 or 5 will have equal segment lengths but different angles between consecutive segments, which will cause the end nodes to be at different distances from the node at level 1. The formal proof is by induction on the level distance. □

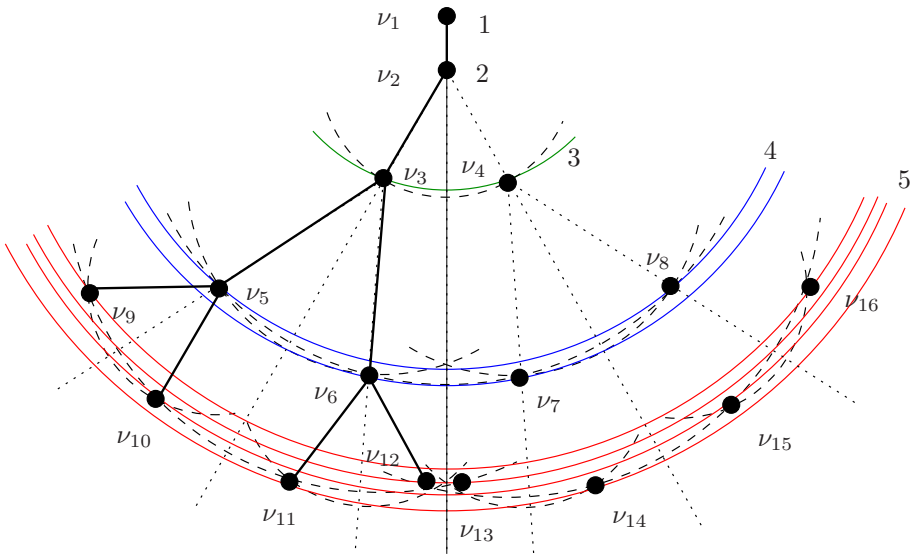

Figure 2: A pruning edge $\{1, 4\}$ prunes either ν_6, ν_7 or ν_5, ν_8.

We now give a basic result on reflections in $\mathbb{R}^K$. For any nonzero vector $y \in \mathbb{R}^K$ let $\mathcal{R}^y$ be the reflection operator through the hyperplane passing through the origin and normal to y.

4.2 Lemma
Let $x \neq y \in \mathbb{R}^K$ *and* $z \in \mathbb{R}^K$ *such that* z *is not in the hyperplanes through the origin and normal to* x, y. *Then* $\mathcal{R}^x\mathcal{R}^y z = \mathcal{R}^{\mathcal{R}^x y}\mathcal{R}^x z$.

Proof. The proof is sketched in Fig. 3 for $\mathbb{R}^2$. By considering the reflection $\mathcal{R}^{\mathcal{R}^x y}$ of the map $\mathcal{R}^y$ through $\mathcal{R}^x$, we get $\|z - \mathcal{R}^y z\| = \|\mathcal{R}^x z - \mathcal{R}^{\mathcal{R}^x y}\mathcal{R}^x z\|$. By reflection through $\mathcal{R}^x$ we get $\|O - z\| = \|O - \mathcal{R}^x z\|$ and $\|O - \mathcal{R}^y z\| = \|O - \mathcal{R}^x\mathcal{R}^y z\|$. By reflection through $\mathcal{R}^y$ we get $\|O - z\| = \|O - \mathcal{R}^y z\|$. By reflection through $\mathcal{R}^{\mathcal{R}^x y}$ we get $\|O - \mathcal{R}^x z\| = \|O - \mathcal{R}^{\mathcal{R}^x y}\mathcal{R}^x z\|$. The triangles $\triangle(z, O, \mathcal{R}^y z)$ and $\triangle(\mathcal{R}^x z, O, \mathcal{R}^{\mathcal{R}^x y}\mathcal{R}^x z)$ are then equal because the side lengths are pairwise equal. Also, reflection of $\triangle(z, O, \mathcal{R}^y z)$ through $\mathcal{R}^x$ yields $\triangle(z, O, \mathcal{R}^y z) = \triangle(\mathcal{R}^x z, O, \mathcal{R}^x\mathcal{R}^y z)$, whence $\mathcal{R}^{\mathcal{R}^x y}\mathcal{R}^x z = \mathcal{R}^x\mathcal{R}^y z$. □

For $v > K$ and $x \in X$ we now define partial reflection operators:

$$g_v(x) = (x_1, \ldots, x_{v-1}, R_x^v(x_v), \ldots, R_x^v(x_n)). \tag{2}$$

The g_v's map an embedding x to its partial reflection with first branch at v. It is easy to show that the g_v's are injective with probability 1 and idempotent.

4.3 Lemma
For $x \in X$ *and* $u, v \in V$ *such that* $u, v > K$, $g_u g_v(x) = g_v g_u(x)$.

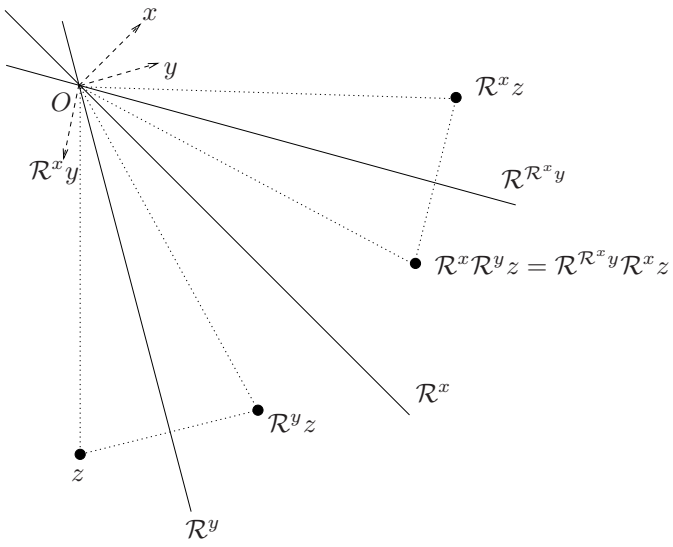

Figure 3: Reflecting through $\mathcal{R}^y$ first and $\mathcal{R}^x$ later is equivalent to reflecting through $\mathcal{R}^x$ first and (the reflection of $\mathcal{R}^y$ through $\mathcal{R}^x$) later.

Proof. Assume without loss of generality $u < v$. Then:

$$\begin{aligned} g_u g_v(x) &= g_u(x_1, \ldots, x_{v-1}, R^v_x(x_v), \ldots, R^v_x(x_n)) \\ &= (x_1 \ldots, x_{u-1}, R^u_{g_v(x)}(x_u), \ldots, R^u_{g_v(x)}R^v_x(x_v), \ldots, R^u_{g_v(x)}R^v_x(x_n)) \\ &= (x_1 \ldots, x_{u-1}, R^u_x(x_u), \ldots, R^v_{g_u(x)}R^u_x(x_v), \ldots, R^v_{g_u(x)}R^u_x(x_n)) \\ &= g_v(x_1, \ldots, x_{u-1}, R^u_x(x_u), \ldots, R^u_x(x_n)) \\ &= g_v g_u(x), \end{aligned}$$

where $R^u_{g_v(x)}R^v_x(x_w) = R^v_{g_u(x)}R^u_x(x_w)$ for each $w \geq v$ by Lemma 4.2. □

We define the *discretization group* to be the group $\mathcal{G}_D = \langle g_v \mid v > K \rangle$ generated by the g_v's.

4.4 Corollary
With probability 1, $\mathcal{G}_D$ is an Abelian group isomorphic to C_2^{n-K}.

For all $v > K$ let $\gamma_v = (1, \ldots, 1, -1_v, \ldots, -1)$ be the vector consisting of one's in the first $v-1$ components and -1 in the last components. Then the g_v actions are naturally mapped onto the chirality functions.

4.5 Lemma
For all $x \in X$, $\chi(g_v(x)) = \chi(x) \odot \gamma_v$, where $\odot$ is the componentwise vector multiplication.

Proof. This follows by definition of g_v and of chirality of an embedding. □

Because, by Alg. 1, each $x \in X$ has a different chirality, for all $x, x' \in X$ there is $g \in \mathcal{G}_D$ such that $x' = g(x)$, i.e. the action of $\mathcal{G}_D$ on X is transitive. By Thm. 4.1, the distances associated to the discretization edges are invariant with respect to the discretization group.

4.2 The pruning group

Consider a pruning edge $\{u, v\} \in E_P$. By Thm. 4.1, with probability 1 we have $d_{uv} \in H^{uv}$, otherwise G cannot be a YES instance (against the hypothesis). Also, again by Thm. 4.1, $d_{uv} = \|x_u - x_v\| \neq \|g_w(x)_u - g_w(x)_v\|$ for all $w \in \{u + K + 1, \ldots, v\}$ (e.g. the distance $\|\nu_1 - \nu_9\|$ in Fig. 2 is different from all its reflections $\|\nu_1 - \nu_h\|$, with $h \in \{10, 11, 12\}$, w.r.t. g_4, g_5). We therefore define the *pruning group*

$$\mathcal{G}_P = \langle g_w \mid w > K \wedge \forall \{u, v\} \in E_P \; (w \notin \{u + K + 1, \ldots, v\}) \rangle.$$

By definition, $\mathcal{G}_P \leq \mathcal{G}_D$ and the distances associated with the pruning edges are invariant with respect to $\mathcal{G}_P$.

4.6 Theorem
The action of $\mathcal{G}_P$ on X is transitive with probability 1.

Proof. Let $x, x' \in X$, we aim to show that $\exists g \in \mathcal{G}_P$ such that $x' = g(x)$ with probability 1. Since the action of $\mathcal{G}_D$ on X is transitive, $\exists g \in \mathcal{G}_D$ with $x' = g(x)$. Now suppose $g \notin \mathcal{G}_P$, then there is a pruning edge $\{u, v\} \in E_P$ and an $\ell \in \mathbb{N}$ s.t. $g = \prod_{h=1}^{\ell} g_{v_h})$ for some vertex set $\{v_1, \ldots, v_\ell > K\}$ including at least one vertex $w \in \{u + K + 1, \ldots, v\}$. By Thm. 4.1, as remarked above, this implies that $d_{uv} = \|x_u - x_v\| \neq \|g_w(x)_u - g_w(x)_v\|$ with probability 1. If the set $Q = \{v_1, \ldots, v_\ell\} \cap \{u + K + 1, \ldots, v\}$ has cardinality 1, then g_w is the only component of g not fixing d_{uv}, and hence $x' = g(x) \notin X$, against the hypothesis. Otherwise, the probability of another $z \in Q \setminus \{w\}$ yielding $\|x_u - x_v\| = \|g_z g_w(x)_u - g_z g_w(x)_v\|$, notwithstanding the fact that $\|g_w(x)_u - g_w(x)_v\| \neq \|x_u - x_v\| \neq \|g_z(x)_u - g_z(x)_v\|$, is zero; and by induction this also covers any cardinality of Q. Therefore $g \in \mathcal{G}_P$ and the result follows. □

The cardinality of X was shown to be a power of two with probability 1 in the unpublished technical report [9]. We provide a shorter and clearer proof.

4.7 Theorem
With probability 1, $\exists \ell \in \mathbb{N} \; |X| = 2^\ell$.

Proof. Since $\mathcal{G}_D \cong C_2^{n-K}$, $|\mathcal{G}_D| = 2^{n-K}$. Since $\mathcal{G}_P \leq \mathcal{G}_D$, $|\mathcal{G}_P|$ divides the order of $|\mathcal{G}_D|$, which implies that there is an integer ℓ with $|\mathcal{G}_P| = 2^\ell$. By Thm. 4.6, the action of $\mathcal{G}_P$ on X only has one orbit, i.e. $\mathcal{G}_P x = X$ for any $x \in X$. By idempotency, for $g, g' \in \mathcal{G}_P$, if $gx = g'x$ then $g = g'$. This implies $|\mathcal{G}_P x| = |\mathcal{G}_P|$. Thus, for any $x \in X$, $|X| = |\mathcal{G}_P x| = |\mathcal{G}_P| = 2^\ell$. □

References

[1] L. Blumenthal. *Theory and Applications of Distance Geometry*. Oxford University Press, Oxford, 1953.

[2] R. Connelly. Generic global rigidity. *Discrete Computational Geometry*, 33:549–563, 2005.

[3] G.M. Crippen and T.F. Havel. *Distance Geometry and Molecular Conformation*. Wiley, New York, 1988.

[4] Q. Dong and Z. Wu. A geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data. *Journal of Global Optimization*, 26:321–333, 2003.

[5] C. Lavor, L. Liberti, N. Maculan, and A. Mucherino. Recent advances on the discretizable molecular distance geometry problem. *European Journal of Operational Research*, in revision (invited survey).

[6] C. Lavor, L. Liberti, N. Maculan, and A. Mucherino. The discretizable molecular distance geometry problem. *Computational Optimization and Applications*, to appear.

[7] C. Lavor, A. Mucherino, L. Liberti, and N. Maculan. On the solution of molecular distance geometry problems with interval data. In *Proceedings of the International Workshop on Computational Proteomics*, Hong Kong, 2010. IEEE.

[8] L. Liberti, C. Lavor, and N. Maculan. A branch-and-prune algorithm for the molecular distance geometry problem. *International Transactions in Operational Research*, 15:1–17, 2008.

[9] L. Liberti, B. Masson, C. Lavor, J. Lee, and A. Mucherino. On the number of solutions of the discretizable molecular distance geometry problem. Technical Report 1010.1834v1[cs.DM], arXiv, 2010.

[10] J.J. Moré and Z. Wu. Global continuation for distance geometry problems. *SIAM Journal of Optimization*, 7(3):814–846, 1997.

[11] A. Mucherino, C. Lavor, and L. Liberti. The discretizable distance geometry problem. *Optimization Letters*, in revision.

[12] J. Richter-Gebert and G. Ziegler. Oriented matroids. In J. Goodman and J. O'Rourke, editors, *Handbook of Discrete and Computational Geometry*, pages 129–151. CRC Press, Boca Raton, 2004.

[13] T. Schlick. *Molecular modelling and simulation: an interdisciplinary guide*. Springer, New York, 2002.