The discretizable molecular distance geometry problem is easier on proteins

Leo Liberti, Carlile Lavor, and Antonio Mucherino

Abstract Distance geometry methods are used to turn a set of interatomic distances given by NMR experiments into a consistent molecular conformation. In a set of papers (see the survey [8]) we proposed a Branch-and-Prune (BP) algorithm for computing the set X of all incongruent embeddings of a given protein backbone. Although BP has a worst-case exponential running time in general, we always noticed a linear-like behaviour in computational experiments. In this paper we provide a theoretical explanation to our observations. We show that the BP is fixed-parameter tractable on protein-like graphs, and empirically show that the parameter is constant on a set of proteins from the Protein Data Bank.

1 Introduction

We consider the following decision problem [9]:

> DISCRETIZABLE MOLECULAR DISTANCE GEOMETRY PROBLEM (DMDGP). Given 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 $v > 3$ and $u \in V$ with $1 \leq v - u \leq 3$, $\{u,v\} \in E$ (DISCRETIZATION)

L. Liberti
LIX, École Polytechnique, 91128 Palaiseau, France
e-mail: liberti@lix.polytechnique.fr

C. Lavor
Dept. of Applied Maths (IMECC-UNICAMP), State Univ. of Campinas, 13081-970, Campinas - SP, Brazil
e-mail: clavor@ime.unicamp.br

A. Mucherino
IRISA, Université de Rennes I, France
e-mail: antonio.mucherino@irisa.fr

2. strict triangular inequalities $d_{v-2,v} < d_{v-2,v-1} + d_{v-1,v}$ hold for all $v > 2$ (NON-COLLINEARITY),

and given an embedding $x' : \{1,2,3\} \to \mathbb{R}^3$, is there an embedding $x : V \to \mathbb{R}^3$ extending x', such that

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

An embedding x on V extends an embedding x' on $U \subseteq V$ if x', as a function, is the restriction of x to U; an embedding is feasible if it satisfies (1). We also consider the following problem variants:

- DMDGP$_K$, i.e. the family of decision problems (parametrized by the positive integer K) obtained by replacing each symbol '3' in the DMDGP definition by the symbol 'K';
- the $^{\mathsf{K}}$DMDGP, where K is given as part of the input (rather than being a fixed constant as in the DMDGP$_K$).

In both variants, strict triangular inequalities are replaced by strict simplex inequalities, see Eq. (11) in [7]. We remark that DMDGP = DMDGP$_3$. Other related problems also exist in the literature, such as the DISCRETIZABLE DISTANCE GEOMETRY PROBLEM (DDGP) [18], where the DISCRETIZATION axiom is relaxed to require that each vertex $v > K$ has at least K adjacent predecessors. The original results in this paper, however, only refer to the DMDGP and its variants.

Statements 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]. We also point out that a statement might hold with probability 1 with respect to a set which has itself Lebesgue measure 0 in a larger set. For example, we will show that the set of $^{\mathsf{K}}$DMDGP instances having an incongruent solution set X with $|X| = 2^\ell$ for some $\ell \in \mathbb{N}$ has measure 1 into the set of all YES instances, which itself is a set of measure 0 in the set of all $^{\mathsf{K}}$DMDGP instances.

The DISCRETIZATION axiom guarantees that the locus of the points that embed 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 with probability 1. The complementary zero-measure set contains instances that do not satisfy the NON-COLLINEARITY axiom, and which might yield loci for v with zero or uncountably many points. We remark that if the intersection of the three spheres is empty, then the instance is a NO one. We solve $^{\mathsf{K}}$DMDGP instances using a recursive algorithm called Branch-and-Prune (BP) [13]: at level v, the search is branched according to the (at most two) 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. The BP has exponential worst-case complexity.

The $^{\mathsf{K}}$DMDGP and its variants are related to the MOLECULAR DISTANCE GEOMETRY PROBLEM (MDGP), i.e. find an embedding in $\mathbb{R}^3$ of a given simple weighted undirected graph. We denote the K-dimensional generalization of the MDGP (with K part of the input) by DISTANCE GEOMETRY PROBLEM (DGP),

and the variant with K fixed by DGP_K. The MDGP is a good model for determining the structure of molecules given a set of inter-atomic distances [14, 11], which are usually given by Nuclear Magnetic Resonance (NMR) experiments [22], a technique which allows the detection of inter-atomic distances below 5.5Å. The DGP has applications to wireless sensor networks [5], statics, robotics and graph drawing among others. In general, the MDGP and DGP implicitly require a search in a continuous Euclidean space [14]. $^{\mathsf{K}}$DMDGP instances describe rigid graphs [6], in particular Henneberg type I graphs [12].

The DMDGP is a model for protein backbones. 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 $d_{v-3,v-2} + d_{v-2,v-1}$ [13].

Over the years, we noticed that the CPU time behaviour of the BP on protein instances looked more linear than exponential. In this paper we give a theoretical motivation for this observation. More precisely, there are cases where BP is actually Fixed-Parameter Tractable (FPT), and we empirically verify on 45 proteins from the PDB [1] that they belong to these cases, and always with the parameter value set to the constant 4. The strategy is as follows: we first show that DMDGP_K is **NP**-hard (Sect. 3), then we show that the number of leaf nodes in the BP search tree is a power of 2 with probability 1 (Sect. 4.2), and finally we use this information to construct a Directed Acyclic Graph (DAG) representing the number of leaf nodes in function of the graph edges (Sect. 5). This DAG allows us to show that the BP is FPT on a class of graphs which provides a good model for proteins (Sect. 5.1).

2 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. Let X ne the set of embeddings (modulo translations and rotations) solving a given $^{\mathsf{K}}$DMDGP instance.

Since vertex v can be placed in at most two possible positions (the intersection of K spheres in $\mathbb{R}^K$), the BP algorithm tests each in turn, and calls itself recursively for every feasible position. BP exploits other edges (than those granted by the DISCRETIZATION axiom) in order to prune 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 $\mathrm{BP}(K+1, x', \emptyset)$ (see Alg. 1), where x' is the initial embedding of the first K vertices mentioned in the $^{\mathsf{K}}$DMDGP definition. By the

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

Require: A vertex $v \in V \setminus [K]$, a partial embedding $\bar{x} = (x_1, \ldots, x_{v-1})$, a set X.
1: $T = \bigcap_{\substack{u \in N(v) \\ u < v}} S_{uv}^{\bar{x}}$;
2: **for** $p \in T$ **do**
3: $x \leftarrow (\bar{x}, p)$
4: **if** $v = n$ **then**
5: $X \leftarrow X \cup \{x\}$
6: **else**
7: BP($v+1$, x, X)
8: **end if**
9: **end for**

$^{\mathsf{K}}$DMDGP axioms, $|T| \leq 2$. At termination, X contains all embeddings (modulo rotations and translations) extending x' [13, 9]. Embeddings $x \in X$ can be represented by sequences $\chi(x) \in \{-1, 1\}^n$ representing left/right choices when traversing a branch from root to leaf of the search tree. More precisely: (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] 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).

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 [8], including, very recently, problems with interval-type uncertainties on some of the distance values [10]. 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. [17]), the performance of the BP algorithm is impressive from the point of view of both efficiency and reliability.

3 Complexity

Any class of YES instances where each vertex v only has distances to the K immediate predecessors provides a full BP binary search tree (after level K), and therefore shows that the BP is an exponential-time algorithm in the worst case. One remarkable feature of the computational experiments conducted on our BP implementation [20] on protein instances is that the exponential-time behaviour of the BP algorithm was never noticed empirically.

Restricting d to only take integer values, the DGP$_1$ is **NP**-complete by reduction from SUBSET-SUM, the DGP$_K$ is (strongly) **NP**-hard by reduction from 3-SAT, and the DGP is (strongly) **NP**-hard by induction on K [21]. Only the DGP$_1$ is known

to be **NP**-complete, because if d takes integer values then the YES-certificate x (the embedding) can be chosen to have integer values too.

The DMDGP is **NP**-hard by reduction from SUBSET-SUM (Thm. 3 in [9]). We generalize this result to the $^{\mathsf{K}}$DMDGP.

Theorem 1. *The DMDGP$_K$ is* **NP***-hard for all $K \geq 2$.*

Proof. Let $a = (a_1, \ldots, a_N)$ be an instance of SUBSET-SUM consisting of positive integers, and define an instance of DMDGP$_K$ where $V = \{0, \ldots, KN\}$, E includes $\{i, i+j\}$ for all $j \in \{1, \ldots, K\}$ and $i \in \{0, \ldots, KN-j\}$, and:

$$\forall i \in \{0, \ldots, KN-1\} \quad d_{i,i+1} = a_{\lfloor i/K \rfloor} \tag{2}$$

$$\forall j \in \{2, \ldots, K\}, i \in \{0, \ldots, KN-j\} \quad d_{i,i+j} = \sqrt{\sum_{\ell=1}^{j} d^2_{i+\ell-1,i+\ell}} \tag{3}$$

$$d_{0,KN} = 0. \tag{4}$$

Let $s \in \{-1,1\}^N$ be a solution of the SUBSET-SUM instance a. We let $x_0 = 0$ and for all $i = K(\ell-1)+j > 0$ we let $x_i = x_{i-1} + s_\ell a_\ell e_j$, where e_j is the vector with a one in component j and zero elsewhere. Because $\sum_{\ell \leq N} s_\ell a_\ell = 0$, if s solves the SUBSET-SUM instance a then, by inspection, x solves the corresponding DMDGP instance (2)-(4). Conversely, let x be an embedding that solves (2)-(4), where we assume without loss of generality that $x_0 = 0$. Then (3) ensures that the line through x_i, x_{i-1} is orthogonal to the line through x_{i-1}, x_{i-2} for all $i > 1$, and again we assume without loss of generality that, for all $j \in \{1, \ldots, K\}$, the lines through x_{j-1}, x_j are parallel to the i-th coordinate axis. Now consider the chirality χ of x: because all distance segments are orthogonal, for each $j \leq K$ the j-th coordinate is given by $x_{KN,j} = \sum_{i \bmod K = j} \chi_i a_{\lfloor i/K \rfloor}$. Since $d_{0,KN} = 0$, for all $j \leq K$ we have $0 = x_{KN,j} = \sum_{\ell \leq N} \chi_{K(\ell-1)+j} a_\ell$, which implies that, for all $j \leq K$, $s^j = (\chi_{K(\ell-1)+j} \mid 1 \leq \ell \leq N)$ is a solution for the SUBSET-SUM instance a. □

Corollary 1. *The $^{\mathsf{K}}$DMDGP is* **NP***-hard.*

Proof. Every specific instance of the $^{\mathsf{K}}$DMDGP specifies a fixed value for K and hence belongs to the DMDGP$_K$. Hence the result follows by inclusion. □

4 Partial reflection symmetries

The results in this section are also found in [16], but the presentation below, which is based on group theory, is new, and (we hope) clearer. 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 $T = \{x_v^0, x_v^1\}$ 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^1 is the reflection of x_v^0 through the hyperplane defined by $x_{v-K}, \ldots, x_{v-1}$. Denote this reflection operator by R_x^v.

Theorem 2 (Cor. 4.6 and Thm. 4.9 in [16]). *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 $\|x_v - x_u\| \neq \|x'_v - x_u\|$.*

We sketch the proof in Fig. 1 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.

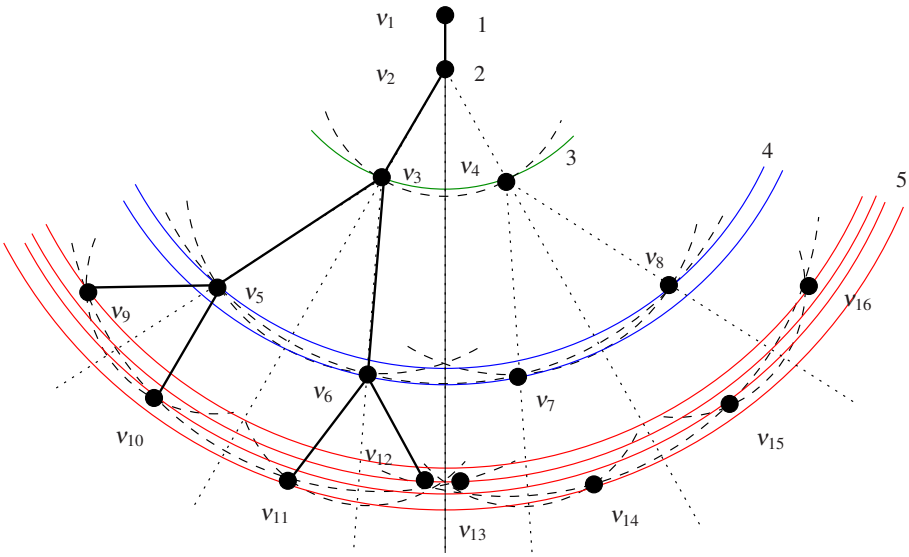

Fig. 1 A pruning edge $\{1,4\}$ prunes either v_6, v_7 or v_5, v_8.

For any nonzero vector $y \in \mathbb{R}^K$ let ρ^y be the reflection operator through the hyperplane passing through the origin and normal to y. If y is normal to the hyperplane defined by $x_{v-K}, \ldots, x_{v-1}$, then $\rho^y = R^v_x$.

Lemma 1. *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 $\rho^x \rho^y z = \rho^{\rho^x y} \rho^x z$.*

Proof. Fig. 2 gives a proof sketch for $K = 2$. By considering the reflection $\rho^{\rho^x y}$ of

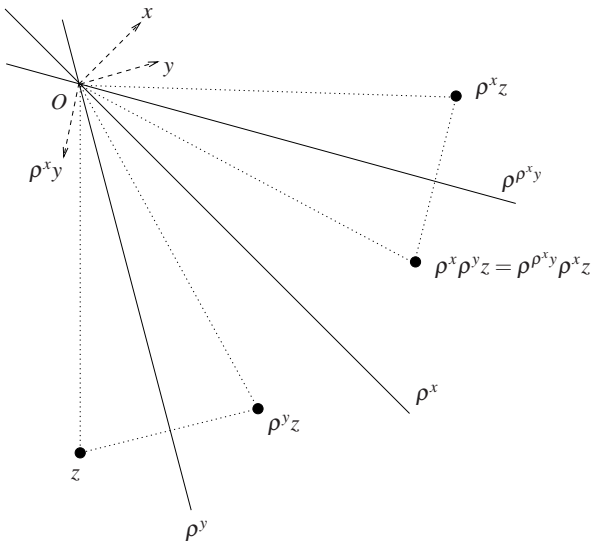

Fig. 2 Reflecting through ρ^y first and ρ^x later is equivalent to reflecting through ρ^x first and (the reflection of ρ^y through ρ^x) later.

the map ρ^y through ρ^x, we get $\|z - \rho^y z\| = \|\rho^x z - \rho^{\rho^x y} \rho^x z\|$. By reflection through ρ^x we get $\|O - z\| = \|O - \rho^x z\|$ and $\|O - \rho^y z\| = \|O - \rho^x \rho^y z\|$. By reflection through ρ^y we get $\|O - z\| = \|O - \rho^y z\|$. By reflection through $\rho^{\rho^x y}$ we get $\|O - \rho^x z\| = \|O - \rho^{\rho^x y} \rho^x z\|$. The triangles $\triangle(z, O, \rho^y z)$ and $\triangle(\rho^x z, O, \rho^{\rho^x y} \rho^x z)$ are then equal because the side lengths are pairwise equal. Also, reflection of $\triangle(z, O, \rho^y z)$ through ρ^x yields $\triangle(z, O, \rho^y z) = \triangle(\rho^x z, O, \rho^x \rho^y z)$, whence $\rho^{\rho^x y} \rho^x z = \rho^x \rho^y z$. □

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

$$g_v(x) = (x_1, \ldots, x_{v-1}, R^v_x(x_v), \ldots, R^v_x(x_n)). \tag{5}$$

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. Further, the g_v's commute.

Lemma 2. *For $x \in X$ and $u, v \in V$ such that $u, v > K$, $g_u g_v(x) = g_v g_u(x)$.*

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 1. □

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

Corollary 2. *With probability 1, $\mathscr{G}_D$ is an Abelian group isomorphic to C_2^{n-K} (the Cartesian product consisting of $n-K$ copies of the cyclic group of order 2).*

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.

Lemma 3. *For all $x \in X$, $\chi(g_v(x)) = \chi(x) \circ \gamma_v$, where $\circ$ is the Hadamard (i.e., component-wise) product.*

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 $g \in \mathscr{G}_D$ such that $x' = g(x)$, i.e. the action of $\mathscr{G}_D$ on X is transitive. By Thm. 2, 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. 2, with probability 1 we have $d_{uv} \in H^{uv}$, otherwise G cannot be a YES instance (against the hypothesis). Also, again by Thm. 2, $d_{uv} = \|x_u - x_v\| \neq \|g_w(x)_u - g_w(x)_v\|$ for all $w \in \{u+K+1, \ldots, v\}$ (e.g. the distance $\|v_1 - v_9\|$ in Fig. 1 is different from all its reflections $\|v_1 - v_h\|$, with $h \in \{10, 11, 12\}$, w.r.t. g_4, g_5). We therefore define the *pruning group*

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

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

Theorem 3. *The action of $\mathscr{G}_P$ on X is transitive with probability 1.*

Proof. This theorem follows from Thm. 5.4 in [15], but here is another, hopefully simpler, proof. Let $x, x' \in X$, we aim to show that $\exists g \in \mathscr{G}_P$ such that $x' = g(x)$ with

probability 1. Since the action of $\mathscr{G}_D$ on X is transitive, $\exists g \in \mathscr{G}_D$ with $x' = g(x)$. Now suppose $g \notin \mathscr{G}_P$, 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. 2, 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 $z \in Q \smallsetminus \{w\}$ 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 $g \in \mathscr{G}_P$ and the result follows. □

Theorem 4. *With probability 1,* $\exists \ell \in \mathbb{N}\ |X| = 2^\ell$.

Proof. Since $\mathscr{G}_D \cong C_2^{n-K}$, $|\mathscr{G}_D| = 2^{n-K}$. Since $\mathscr{G}_P \leq \mathscr{G}_D$, $|\mathscr{G}_P|$ divides the order of $|\mathscr{G}_D|$, which implies that there is an integer ℓ with $|\mathscr{G}_P| = 2^\ell$. By Thm. 3, the action of $\mathscr{G}_P$ on X only has one orbit, i.e. $\mathscr{G}_P x = X$ for any $x \in X$. By idempotency, for $g, g' \in \mathscr{G}_P$, if $gx = g'x$ then $g = g'$. This implies $|\mathscr{G}_P x| = |\mathscr{G}_P|$. Thus, for any $x \in X$, $|X| = |\mathscr{G}_P x| = |\mathscr{G}_P| = 2^\ell$. □

5 Bounded treewidth

The results of the previous section allow us to express the the number of nodes at each level of the BP search tree in function of the level rank and the pruning edges. Fig. 3 shows a DAG $\mathscr{D}_{uv}$ that represents the number of valid BP search tree nodes in function of pruning edges between two vertices $u, v \in V$ such that $v > K$ and $u < v - K$. The first line shows different values for the rank of v w.r.t. u; an arc labelled with an integer i implies the existence of a pruning edge $\{u+i, v\}$ (arcs with $\vee$-expressions replace parallel arcs with different labels). An arc is unlabelled if there is no pruning edge $\{w, v\}$ for any $w \in \{u,\ldots,v-K-1\}$. The vertices of the DAG are arranged vertically by BP search tree level, and are labelled with the number of BP nodes at a given level, which is always a power of two by Thm. 4. A path in this DAG represents the set of pruning edges between u and v, and its incident vertices show the number of valid nodes at the corresponding levels. For example, following unlabelled arcs corresponds to no pruning edge between u and v and leads to a full binary BP search tree with 2^{v-K} nodes at level v.

5.1 Fixed-parameter tractable behaviour

For a given G_D, each possible pruning edge set E_P corresponds to a path spanning all columns in $\mathscr{D}_{1n}$. Instances with diagonal (Prop. 1) or below-diagonal (Prop. 2) E_P paths yield BP trees whose width is bounded by $O(2^{v_0})$ where v_0 is small w.r.t. n.

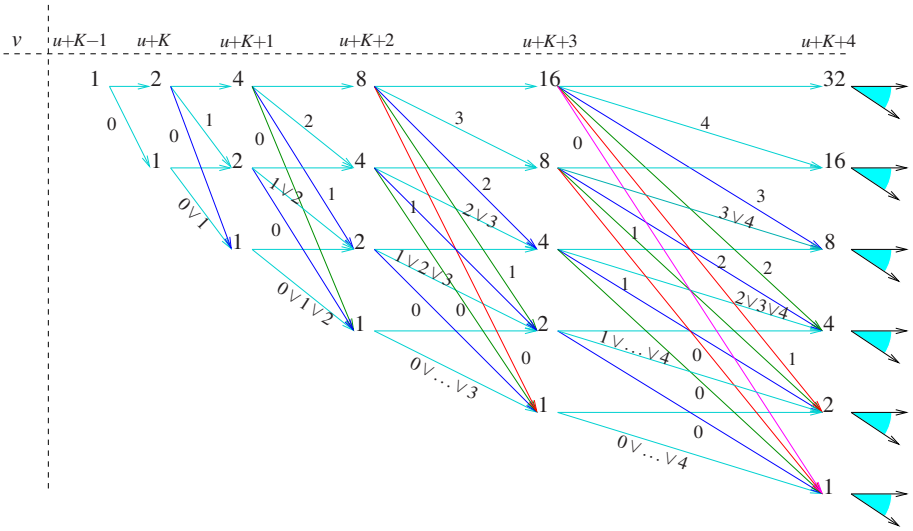

Fig. 3 Number of valid BP nodes (vertex label) at level $u+K+\ell$ (column) in function of the pruning edges (path spanning all columns).

Proposition 1. *If* $\exists v_0 > K$ *s.t.* $\forall v > v_0\ \exists u < v-K$ *with* $\{u,v\} \in E_P$ *then the BP search tree width is bounded by* 2^{v_0-K}.

This corresponds to a path $\mathsf{p}_0 = (1,2,\ldots,2^{v_0-K},\ldots,2^{v_0-K})$ that follows unlabelled arcs up to level v_0 and then arcs labelled v_0-K-1, $v_0-K-1 \vee v_0-K$, and so on, leading to nodes that are all labelled with 2^{v_0-K} (Fig. 4, left).

Proposition 2. *If* $\exists v_0 > K$ *such that every subsequence* s *of consecutive vertices* $>v_0$ *with no incident pruning edge is preceded by a vertex* v_s *such that* $\exists u_s < v_s\ (v_s - u_s \geq |s| \wedge \{u_s, v_s\} \in E_P)$, *then the BP search tree width is bounded by* 2^{v_0-K}.

This situation corresponds to a below-diagonal path, Fig. 4 (right). In general, for those instances for which the BP search tree width has a $O(2^{v_0}\log n)$ bound, the BP has a worst-case running time $O(2^{v_0}L2^{\log n}) = O(Ln)$, where L is the complexity of computing T. Since L is typically constant in n [4], for such cases the BP runs in time $O(2^{v_0}n)$. Let $V' = \{v \in V \mid \exists \ell \in \mathbb{N}\ (v = 2^\ell)\}$.

Proposition 3. *If* $\exists v_0 > K$ *s.t. for all* $v \in V \smallsetminus V'$ *with* $v > v_0$ *there is* $u < v-K$ *with* $\{u,v\} \in E_P$ *then the BP search tree width at level* n *is bounded by* $2^{v_0}n$.

This corresponds to a path along the diagonal 2^{v_0} apart from logarithmically many vertices in V (those in V'), at which levels the BP doubles the number of search nodes (Fig. 5). For a pruning edge set E_P as in Prop. 3, or yielding a path below it, the BP runs in $O(2^{v_0}n^2)$.

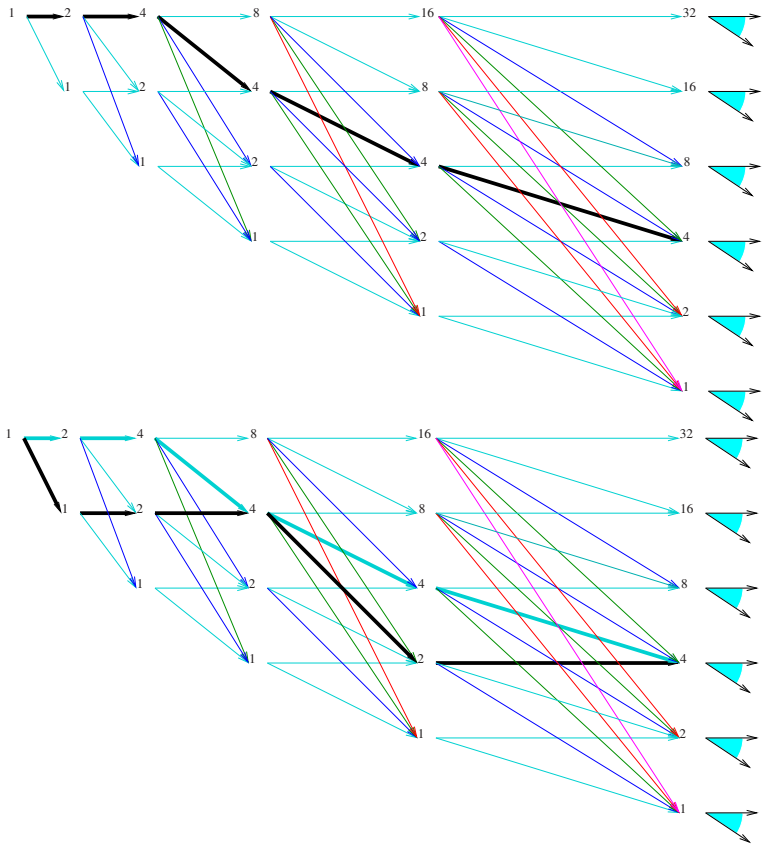

Fig. 4 A path p_0 yielding treewidth 4 (above) and another path below p_0 (below).

5.2 Empirical verification

We consider a set of 45 protein instances from the Protein Data Bank (PDB). Since PDB data include the Euclidean embedding of the protein, we compute all distances, then filter all those that are longer than 5.5Å, which is a reasonable estimate for NMR measures [22]. We then form the weighted graph and embed it with our implementation of the BP algorithm [20]. It turns out that 40 proteins satisfy Prop. 1, and 5 proteins satisfy Prop. 2, all with $v_0 = 4$ (see Table 1). This is consistent with the computational insight [9] that BP empirically displays a polynomial (specifically, linear) complexity on real proteins.

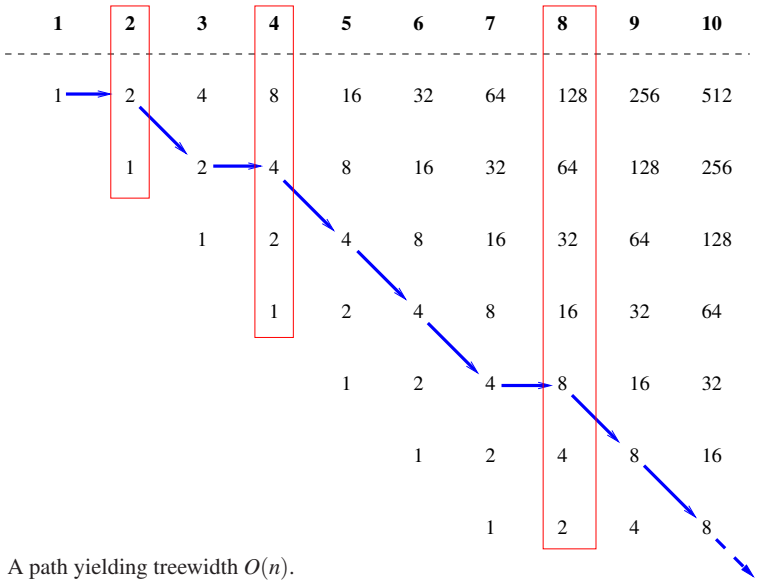

Fig. 5 A path yielding treewidth $O(n)$.

PDB ID	$\|V\|$	v_0
		Prop. 1
1brv	57	4
1a11	75	4
1acw	87	4
1ppt	108	4
1bbl	111	4
1erp	114	4
1aqr	120	4
1k1v	123	4
1h1j	132	4
1ed7	135	4
1dv0	135	4
1crn	138	4
1jkz	138	4
1ahl	147	4
1ptq	150	4
1brz	159	4
1ccq	180	4
1hoe	222	4
1bqx	231	4
1pht	249	4
1a2s	267	4
1jk2	270	4

PDB ID	$\|V\|$	v_0	
		Prop. 1	Prop. 2
1a70	291	269	4
1ag4	309	4	-
2hsy	312	4	-
1acz	324	4	-
1poa	354	4	-
1fs3	372	4	-
1itm	390	4	-
1mbn	459	369	4
1ngl	537	4	-
1b4c	552	280	4
1la3	564	4	-
1a23	567	4	-
1oy2	573	4	-
2ron	726	4	-
1d8v	789	4	-
1rgs	792	203	4
1q8k	900	4	-
1ezo	1110	4	-
1m40	1224	4	-
1bpm	1443	1319	4
1n4w	1610	4	-
1mqq	2032	4	-
3b34	2790	4	-

Table 1 Computation of minimum v_0 in PDB instances.

6 Conclusion

In this paper we provide a theoretical basis to the empirical observation that the BP never seems to attain its exponential worst-case time bound on DMDGP in-

stances from proteins. Other original contributions include a generalization of an **NP**-hardness proof to the ^{K}DMDGP, and a new presentation, based on group theory and involving new proofs, of the fact that the cardinality of the solution set of YES instances of the ^{K}DMDGP is a power of two with probability 1.

References

1. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., Bourne, P.: The protein data bank. Nucleic Acid Research **28**, 235–242 (2000)
2. Connelly, R.: Generic global rigidity. Discrete Computational Geometry **33**, 549–563 (2005)
3. Crippen, G., Havel, T.: Distance Geometry and Molecular Conformation. Wiley, New York (1988)
4. Dong, Q., Wu, Z.: 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. Eren, T., Goldenberg, D., Whiteley, W., Yang, Y., Morse, A., Anderson, B., Belhumeur, P.: Rigidity, computation, and randomization in network localization. IEEE Infocom Proceedings pp. 2673–2684 (2004)
6. Graver, J., Servatius, B., Servatius, H.: Combinatorial Rigidity. American Mathematical Society (1993)
7. Lavor, C., Lee, J., John, A.L.S., Liberti, L., Mucherino, A., Sviridenko, M.: Discretization orders for distance geometry problems. Optimization Letters (DOI: 10.1007/s11590-011-0302-6). DOI 10.1007/s11590-011-0302-6
8. Lavor, C., Liberti, L., Maculan, N., Mucherino, A.: Recent advances on the discretizable molecular distance geometry problem. European Journal of Operational Research **219**, 698–706 (2012)
9. Lavor, C., Liberti, L., Maculan, N., Mucherino, A.: The discretizable molecular distance geometry problem. Computational Optimization and Applications (DOI: 10.1007/s10589-011-9402-6). DOI 10.1007/s10589-011-9402-6
10. Lavor, C., Mucherino, A., Liberti, L., Maculan, N.: On the solution of molecular distance geometry problems with interval data. In: Proceedings of the International Workshop on Computational Proteomics. IEEE, Hong Kong (2010)
11. Lavor, C., Mucherino, A., Liberti, L., Maculan, N.: On the computation of protein backbones by using artificial backbones of hydrogens. Journal of Global Optimization **50**, 329–344 (2011)
12. Liberti, L., Lavor, C.: On a relationship between graph realizability and distance matrix completion. In: V. Kostoglou, G. Arabatzis, L. Karamitopoulos (eds.) Proceedings of BALCOR, vol. I, pp. 2–9. Hellenic OR Society, Thessaloniki (2011)
13. Liberti, L., Lavor, C., Maculan, N.: A branch-and-prune algorithm for the molecular distance geometry problem. International Transactions in Operational Research **15**, 1–17 (2008)
14. Liberti, L., Lavor, C., Mucherino, A., Maculan, N.: Molecular distance geometry methods: from continuous to discrete. International Transactions in Operational Research **18**, 33–51 (2010)
15. Liberti, L., Masson, B., Lavor, C., Lee, J., Mucherino, A.: On the number of solutions of the discretizable molecular distance geometry problem. Tech. Rep. 1010.1834v1[cs.DM], arXiv (2010)
16. Liberti, L., Masson, B., Lee, J., Lavor, C., Mucherino, A.: On the number of solutions of the discretizable molecular distance geometry problem. In: Combinatorial Optimization, Constraints and Applications (COCOA11), *LNCS*, vol. 6831, pp. 322–342. Springer, New York (2011)
17. Moré, J., Wu, Z.: Global continuation for distance geometry problems. SIAM Journal of Optimization **7**(3), 814–846 (1997)

18. Mucherino, A., Lavor, C., Liberti, L.: The discretizable distance geometry problem. Optimization Letters (DOI:10.1007/s11590-011-0358-3)
19. Mucherino, A., Lavor, C., Liberti, L., Talbi, E.G.: A parallel version of the branch & prune algorithm for the molecular distance geometry problem. In: ACS/IEEE International Conference on Computer Systems and Applications (AICCSA10). IEEE conference proceedings, Hammamet, Tunisia (2010)
20. Mucherino, A., Liberti, L., Lavor, C.: `MD-jeep`: an implementation of a branch-and-prune algorithm for distance geometry problems. In: K. Fukuda, J. van der Hoeven, M. Joswig, N. Takayama (eds.) Mathematical Software, *LNCS*, vol. 6327, pp. 186–197. Springer, New York (2010)
21. Saxe, J.: Embeddability of weighted graphs in k-space is strongly NP-hard. Proceedings of 17th Allerton Conference in Communications, Control and Computing pp. 480–489 (1979)
22. Schlick, T.: Molecular modelling and simulation: an interdisciplinary guide. Springer, New York (2002)