An exponential algorithm for the Discretizable Molecular Distance Geometry Problem is polynomial on proteins

Leo Liberti[1], Carlile Lavor[2], Antonio Mucherino[3]

[1] LIX, École Polytechnique, 91128 Palaiseau, France
liberti@lix.polytechnique.fr
[2] Dept. of Applied Maths (IME-UNICAMP), State Univ. of Campinas, 13081-970, Campinas - SP, Brazil clavor@ime.unicamp.br
[3] CERFACS, Toulouse, France antonio.mucherino@cerfacs.fr

Abstract. An important application of distance geometry to biochemistry studies the embeddings of the vertices of a weighted graph in the three-dimensional Euclidean space such that the edge weights are equal to the Euclidean distances between corresponding point pairs. When the graph represents the backbone of a protein, one can exploit the natural vertex order to show that the search space for feasible embeddings is discrete. The corresponding decision problem can be solved using a binary tree based search procedure which is exponential in the worst case. We discuss assumptions that bound the search tree width to a polynomial size, and show empirically that they apply to proteins.
Keywords: Branch-and-Prune, symmetry, distance geometry.

1 Introduction

The Molecular Distance Geometry Problem, which asks to find the embedding in $\mathbb{R}^3$ of a given weighted undirected graph, is a good model for determining the structure of proteins given a set of inter-atomic distances [2]. Its generalization to $\mathbb{R}^K$ is called Distance Geometry Problem (DGP). In general, the MDGP and DGP implicitly require a search in a continuous Euclidean space. Proteins, however, have further structural properties that can be exploited to define subclasses of instances of the MDGP and DGP whose solution set is finite [1]. These instances can be solved with an algorithmic framework called Branch-and-Prune (BP) [1]: this is an iterative algorithm where the i-th atom of the protein can be embedded in $\mathbb{R}^3$ using distances to at least three preceding atoms. Since the intersection of three 3D spheres contains in general two points, the BP gives rise to a binary search tree. In the worst case, the BP is an exponential time algorithm, which is fitting because the MDGP and DGP are **NP**-hard [Saxe, 1979]. Compared to continuous search algorithms, the performance of the BP algorithm is impressive from the point of view of both efficiency and reliability. In this paper we show that the BP has a polynomial worst-case under assumptions found in proteins.

2 Discretizable instances and the BP algorithm

Notation

For all integers $n > 0$, we let $[n] = \{1, \ldots, n\}$. Given an undirected graph $G = (V, E)$ with $|V| = n$, 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. Given a positive integer K, an *embedding* of G in $\mathbb{R}^K$ is a function $x : V \to \mathbb{R}^K$. If $d : E \to \mathbb{R}_+$ is a given edge weight function on $G = (V, E, d)$, an embedding is *valid* for G if $\forall \{u, v\} \in E \; \|x_u - x_v\| = d_{uv}$. For any $U \subseteq V$, an embedding of $G[U]$ (i.e. the subgraph of G induced by U) is a *partial embedding* of G. If x is a partial embedding of G and y is an embedding of G such that $\forall u \in U \, (x_u = y_u)$ then y is an *extension* of x. For a total order $<$ on V and for each $v \in V$, let $\rho(v) = |\{u \in V \mid u \leq v\}|$ be the *rank* of v in V with respect to $<$. The rank is a bijection between V and $[n]$, so we can identify v with its rank and extend arithmetic notation to V so that for $i \in \mathbb{Z}$, $v + i$ denotes the vertex $u \in V$ with $\rho(u) = \rho(v) + i$. For all $v \in V$ and $\ell < \rho(v)$ we denote by $\gamma_\ell(v)$ the set of ℓ immediate predecessors of v. If $U \subseteq V$ with $|U| = h$ such that $G[U]$ is a clique, let $D'(U)$ be the symmetric matrix whose (u, v)-th component is d_{uv}^2 for $u, v \in U$, and let $D(U)$ be $D'(U)$ bordered by a left $(0, 1, \ldots, 1)^\top$ column and a top $(0, 1 \ldots, 1)$ row (both of size $h + 1$). Then the Cayley-Menger formula states that the volume in $\mathbb{R}^{h-1}$ of the h-simplex defined by $G[U]$ is given by $\Delta_{h-1}(U) = \sqrt{\frac{(-1)^h}{2^{h-1}((h-1)!)^2}|D(U)|}$.

Generalized Discretizable Molecular Distance Geometry Problem ($^{\mathsf{K}}$DMDGP). Given an integer $K > 0$, a weighted undirected graph $G = (V, E, d)$ with $d : E \to \mathbb{Q}_+$, a total order $<$ on V and an embedding $x' : [K] \to \mathbb{R}^K$ such that:

1. x' is a valid partial embedding of $G[[K]]$ (Start)
2. G contains all $(K + 1)$-cliques of $<$-consecutive vertices as induced subgraphs (Discretization)
3. $\forall v \in V$ with $v > K$, $\Delta_{K-1}(\gamma_K(v)) > 0$ (Strict Simplex Inequalities),

is there a valid embedding x of G in $\mathbb{R}^K$ extending x'?

We denote by X the set of embeddings solving a $^{\mathsf{K}}$DMDGP instance; X is a finite set [1]. The $^{\mathsf{K}}$DMDGP is **NP**-hard by reduction from the DMDGP [1]. For a partial embedding x of G and $\{u, v\} \in E$ let S_{uv}^x be the sphere centered at x_u with radius d_{uv}. The BP algorithm, used for solving the $^{\mathsf{K}}$DMDGP and its

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

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

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

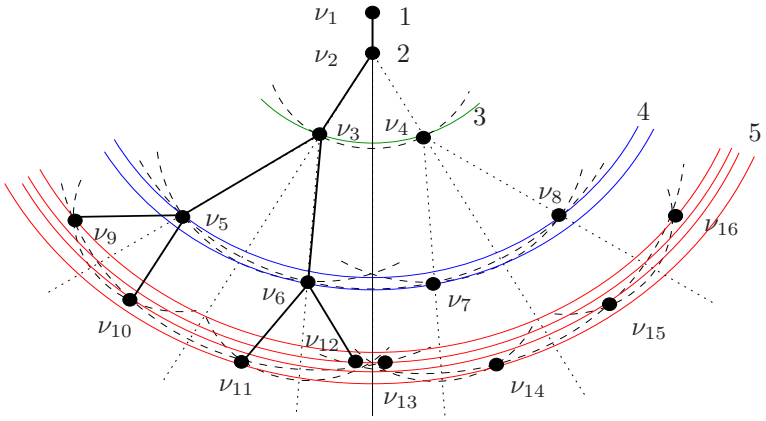
2: $\forall p \in P \; (\, (x \leftarrow (\bar{x}, p))$; **if** $(\rho(v) = n) \; X \leftarrow X \cup \{x\}$ **else** BP($v + 1$, x, X)).

restrictions, is BP($K+1$, x', $\emptyset$) (see Alg. 1). By Strict Simplex Inequalities, $|P| \leq 2$. At termination, X contains all embeddings extending x' [1].

3 BP tree geometry

Since the definition of the $^{\mathsf{K}}$DMDGP requires G to have at least those edges used to satisfy the DISCRETIZATION axiom, we partition E into the sets $E_D = \{\{u,v\} \mid |\rho(v) - \rho(u)| \leq K\}$ and $E_P = E \setminus E_D$. With a slight abuse of notation we call E_D the *discretization distances* (guaranteeing that a DGP instance is in $^{\mathsf{K}}$DMDGP) and E_P the *pruning distances* (used to reduce the search space by pruning the BP tree). Pruning distances might make the set P in Alg. 1 empty or a singleton.

Let G be a YES instance of the $^{\mathsf{K}}$DMDGP, $G_D = (V, E_D, d)$ and let X_D be the set of embeddings of G_D; since G_D has no pruning distances, the BP search tree for G_D is a full binary tree and $|X_D| = 2^{n-K}$. The discretization distances arrange the embeddings so that, at level ℓ, there are $2^{\ell-K}$ possible embeddings x_v for the vertex v with rank ℓ. Furthermore, when $P = \{x_v, x'_v\}$ and the discretization distances to v only involve the K immediate predecessors of v, we have that $x'_v = R^v_x(x_v)$ [3], the reflection of x_v w.r.t. the hyperplane through $x_{v-K}, \ldots, x_{v-1}$. This also implies that the partial embeddings encoded in two BP subtrees rooted at reflected nodes ν, ν' are reflections of each other. This situation is shown in the picture below.

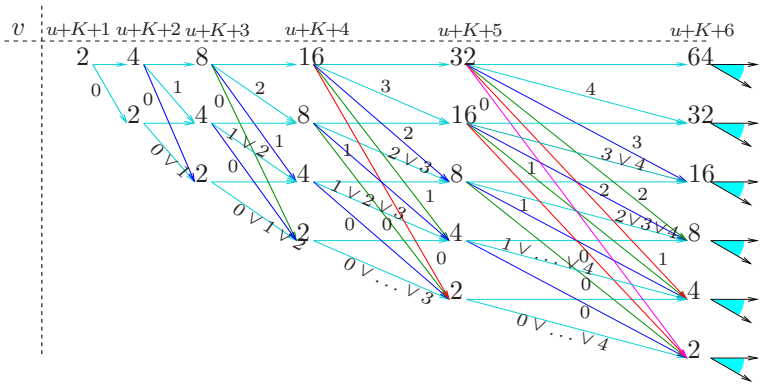


More precisely, with probability 1 we have $\forall v > K, u < v - K \; \exists \; H^{uv} \subseteq \mathbb{R}$ s.t. $|H^{uv}| = 2^{v-u-K}$ and $\forall x \in X \; \|x_v - x_u\| \in H^{uv}$; also $\forall x \in X \; \|x_v - x_u\| = \|R^{u+K}_x(x_v) - x_u\|$ and $\forall x' \in X \; (x'_v \notin \{x_v, R^{u+K}_x(x_v)\} \rightarrow \|x_v - x_u\| \neq \|x'_v - x_u\|)$.

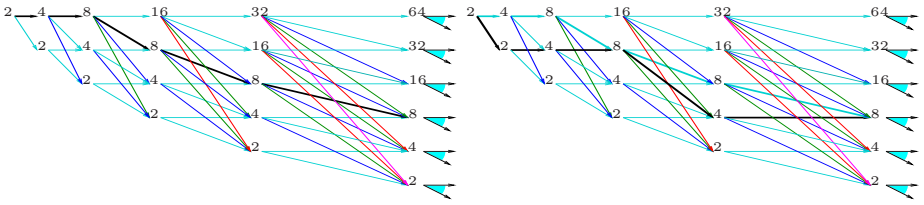
4 BP search trees with bounded width

Consider the BP tree for G_D and assume that there is a pruning distance $\{u,v\} \in E_P$; at level u there are $\max(2^{u-K}, 1)$ nodes, each of which is the root of a subtree with $2^{v-\max(u,K)}$ nodes at level v. By the above remarks, for each such subtree only two nodes will encode a valid embedding for v (we call such nodes *valid*). Thus the number of valid nodes at level $v > K$ is $2^{\max(u-K+1,1)}$.

Consider the following Directed Acyclic Graph (DAG) $\mathcal{D}_{uv}$, used to compute the number of BP nodes in function of pruning distances $\{u,v\}$ with $u < v - K$.



Nodes, arranged vertically, show the number of BP nodes in function of the rank of v w.r.t. u (first line). An arc is labelled with $i_1, \ldots, i_h$ if one of $\{u + i_j, v\}$ (for $j \leq h$) is a pruning distance, and is unlabelled if no such pruning distance exists. A path p in this DAG represents the set of pruning distances between u and v: each node p_ℓ in this path shows the number of valid nodes in the BP search tree at level ℓ. For example, following unlabelled arcs corresponds to no pruning distance between u and v and leads to a full binary BP search tree with 2^{v-K} nodes at level v. Each set of pruning distances E_P corresponds to a longest path in $\mathcal{D}_{1n}$. BP trees have bounded width when these paths are below a diagonal with constant node labels. For example, if $\exists v_0 \in V \smallsetminus [K]$ 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 situation is pictured below (left). Another polynomial class of cases is shown on the right.



Out of a set of 16 protein instances from the Protein Data Bank (PDB), all yield BP trees of bounded width (with $v_0 = 4$). This empirically illustrates the polynomiality of BP on real proteins.

References

1. C. Lavor, L. Liberti, N. Maculan, and A. Mucherino. The discretizable molecular distance geometry problem. *Comp. Opt. Appl.*, to appear.
2. L. Liberti, C. Lavor, A. Mucherino, and N. Maculan. Molecular distance geometry methods: from continuous to discrete. *Int. Trans. Op. Res.*, 18:33–51, 2010.
3. L. Liberti, B. Masson, C. Lavor, J. Lee, and A. Mucherino. Technical Report 1010.1834v1[cs.DM], arXiv, 2010.