Molecular distance geometry methods: From continuous to discrete

LEO LIBERTI[1], CARLILE LAVOR[2], ANTONIO MUCHERINO[1], NELSON MACULAN[3]

[1] *LIX, École Polytechnique, F-91128 Palaiseau, France*
Email: liberti@lix.polytechnique.fr; mucherino@lix.polytechnique.fr

[2] *Department of Applied Mathematics (IMECC-UNICAMP), State University of Campinas, C.P. 6065, 13081-970, Campinas - SP, Brazil*
Email: clavor@ime.unicamp.br

[3] *Federal University of Rio de Janeiro (COPPE–UFRJ), C.P. 68511, 21945-970, Rio de Janeiro - RJ, Brazil*
Email: maculan@cos.ufrj.br

November 14, 2009

Abstract

Distance geometry problems arise from the need to position entities in the Euclidean K-space given some of their respective distances. Entities may be atoms (molecular distance geometry), wireless sensors (sensor network localization), or abstract vertices of a graph (graph drawing). In the context of molecular distance geometry, the distances are usually known because of chemical properties and Nuclear Magnetic Resonance experiments; sensor networks can estimate their relative distance by recording the power loss during a two-way exchange; finally, when drawing graphs in 2D or 3D, the graph to be drawn is given, and therefore distances between vertices can be computed. Distance geometry problems involve a search in a continuous Euclidean space, but sometimes the problem structure helps reduce the search to a discrete set of points. In this paper we survey some continuous and discrete methods for solving some problems of molecular distance geometry.

1 Introduction

In several situations, when placing entities in the Euclidean space, some inter-entity distances are known in advance, and the chosen positions for the entities must respect the known distances. Problems of this type are called Distance Geometry Problems (DGPs). The three main realms of application of DGPs are:

- the Molecular Distance Geometry Problem (MDGP) and its variants;
- the (wireless) Sensor Network Localization Problem (SNLP);
- Graph Drawing (GD).

In the MDGP the entities are atoms, sets of entities are molecules and a subset of inter-atomic distances may be known because of the type of chemical bonds between atoms, or by means of Nuclear Magnetic Resonance (NMR) experiments. Typically, the MDGP involves placing atoms in $\mathbb{R}^3$ [9, 19, 39, 27, 33]. In the SNLP the entities are wireless sensors: a pair of sensors can estimate their distance by measuring the quantity of battery power necessary to a two-way communication. Usually, the entities are embedded in $\mathbb{R}^2$. One particular property of the SNLP, which distinguishes it from the MDGP and GD, is that sensor networks almost always have a subset of fixed sensor whose position is known in advance (these are called *anchors*) [13, 3, 51]. GD is the problem of deriving representations in the plane or in the three-dimensional space of graphs, with the aim of finding convenient visualizations of certain properties of the graph. Basically, it consists in finding an embedding in $\mathbb{R}^K$ of a given weighted graph $G = (V, E, d)$ where $d : E \to \mathbb{R}_+$ (the GD community contributes a well-established series of symposia with published proceedings in the Lecture Notes in Computer Science Springer series, see `www.graphdrawing.org`).

Although DGPs are essentially a search in a continuous Euclidean space, sometimes the topology of the graph supporting the known distances guarantees sufficient structure for the search to become discrete. In this paper we focus on the MDGP and its variants. We present a survey of some continuous and discrete methods, highlighting the role of the other communities (mostly SNLP) in the development of the "discretization" of MDGP solution methods.

This survey consists in two main sections. In Section 2 we formalize the MDGP, and introduce some continuous formulations and solution methods. In Section 3 we explain under what conditions the search only need span a discrete set; we then describe a discrete MDGP variant and discuss an algorithmic framework, called Branch-and-Prune (BP), that solves it.

2 Search in continuous space

In this section we review some continuous formulations and methods used to solve the MDGP for general molecules, when no further structure on the underlying graph topology is known.

2.1 Continuous formulations

A molecule is represented by a weighted undirected graph $G = (V, E, d)$ where V is the set of atoms, E is a symmetric relation on V whose members connect atoms whose distances are known through either chemical bond analysis or NMR experiments, and d is a function $E \to \mathbb{R}_+$ where d_{ij} is the Euclidean distance between atom i and atom j (for all $\{i, j\} \in E$). Given an integer $K > 0$, we wish to determine a function $x : V \to \mathbb{R}^K$ (called *embedding*) in such a way that:

$$\forall \{i, j\} \in E \quad ||x_i - x_j|| = d_{ij}, \tag{1}$$

where the norm $|| \cdot ||$ is taken to be the Euclidean norm. We remark that Eq. (1) is a system of nonlinear equations.

Let $\bar{X} = \{x \mid x \text{ satisfies (1)}\}$. If no atom is fixed to a particular position, then $\bar{X}$ is either empty or uncountable, because for any solution $\bar{x} \in \bar{X}$ and any orthogonal transformation T of $\mathbb{R}^K$, we have $T\bar{x} \in \bar{X}$ by definition of orthogonality — and there is an uncountable number of orthogonal transformations of $\mathbb{R}^3$. We define an equivalence relation $\sim$ on $\bar{X}$ given by $\bar{x} \sim \bar{y}$ if and only if there exists an orthogonal transformation T such that $\bar{y} = T\bar{x}$, and let $X = \bar{X}/\sim$. Then a natural injection of X into $\mathbb{R}^K$ can be obtained by fixing the positions of K atoms in V such that they satisfy (1). In general, when no structure on G is known a priori, X may be empty, finite or again uncountable.

Numerically, solving the nonlinear equations (1) directly is very difficult. Some subsystems of (1), however, are sometimes considered and solved as part of more complex specific solution methods [11, 10]. More often, system (1) is re-cast as a penalty function to be minimized:

$$\min_x \sum_{\{i,j\} \in E} (||x_i - x_j||^2 - d_{ij}^2)^2. \tag{2}$$

In the above, the left and right hand sides of (1) are squared before taking their square difference, so that the resulting nonlinear objective function does not involve a square root operation (which may pose some numerical issues for arguments close to zero). We remark that although a sum of squares, (2) is actually a nonconvex optimization problem in x, and falls into the category of Global Optimization (GO).

The decision problem associated with system (1) is the following:

> Molecular Distance Geometry Problem. Given an integer $K > 0$ and a weighted undirected graph $G = (V, E, d)$ with $d : E \to \mathbb{R}_+$ find $x : V \to \mathbb{R}^K$ such that (1) holds.

The MDGP is strongly **NP**-complete for $K = 1$ and strongly **NP**-hard for $K > 1$ [48].

Since NMR measurements are subject to noise, (1) is sometimes replaced by a system of nonlinear inequalities:

$$\forall \{i, j\} \in E \quad d^L_{ij} \le ||x_i - x_j|| \le d^U_{ij}. \tag{3}$$

The corresponding decision problem is the inexact Molecular Distance Geometry Problem (iMDGP), defined as the MDGP with (1) replaced by (3). A formulation for the iMDGP is as follows:

$$\min_x \sum_{\{i,j\} \in E} \left[\left(\max\left((d^L_{ij})^2 - ||x_i - x_j||^2, 0\right)\right)^2 + \left(\max\left(||x_i - x_j||^2 - (d^U_{ij})^2, 0\right)\right)^2 \right]. \tag{4}$$

Usually, though, the terms are weighted as follows:

$$\min_x \sum_{\{i,j\} \in E} \left[\left(\max\left(\frac{(d^L_{ij})^2 - ||x_i - x_j||^2}{(d^L_{ij})^2}, 0\right)\right)^2 + \left(\max\left(\frac{||x_i - x_j||^2 - (d^U_{ij})^2}{(d^U_{ij})^2}, 0\right)\right)^2 \right]. \tag{5}$$

2.2 Continuous methods

In this section we review some MDGP solution methods based on searching the continuous Euclidean space $\mathbb{R}^3$.

2.2.1 General-purpose methods

Given a nonconvex continuous GO problem such as (2), the first — and obvious — idea is to attempt to solve it with existing general-purpose GO methods. In [26], comparative computational results on some artificial MDGP instances from [39, 25] are obtained using the following GO methods [30].

- The spatial Branch-and-Bound (sBB) algorithm is an enumerative tree-like search in continuous space [30, 37] based on partitioning the original feasible region into rectangular boxes. Upper and lower bounds are derived for the subproblem defined on each box; if the bounds are closer than a given ε tolerance, the global optimum for the box is deemed found and the box is discarded, otherwise the box is recursively partitioned. The search terminates when all boxes have been explored. In general, upper bounds are given by any local optimum of (2), whilst lower bounds are given by a suitable convex relaxation of the problem [50, 52, 4]. In the present case, however, we know that if a solution of (1) exists, then the objective function value of (2) is exactly zero, so zero is the tightest possible lower bound at the root node of the sBB search. The sBB is essentially an ε-approximation algorithm for solving nonconvex Nonlinear Programs (NLPs) and Mixed-Integer Nonlinear Programs (MINLPs). Since searches in nonlinear manifolds of a continuous space can rarely find exact optima, the best one can do in such cases is ε-approximation; this is way the sBB is usually referred to as an *exact GO method*.
- A GO variant of the Variable Neighbourhood Search (VNS) [17] metaheuristic, initially inspired by [38], first described in [31] and recently extended to deal with constrained nonconvex MINLPs [36], is a stochastic algorithm based on iteratively exploring increasingly larger neighbourhoods of the current best optimum (the *incumbent*) until either a new, better optimum is found or a termination condition stops the search.

- The multistart-like heuristic algorithm SobolOpt [24] is a stochastic algorithm that employs Sobol' pseudo-random sequences to draw points from which to start local descents. Such sequences guarantee a good spatial distribution that keeps holding true in a set of projected subspaces.

Whereas the sBB has an optimality guarantee, the two stochastic algorithms only carry a guarantee in probability in infinite time. Thus, due to the usual trade-off between effort and results, heuristics are expected to perform better, CPU time-wise, than the exact algorithm. It therefore comes to somewhat of a surprise that the computational results reported in [26] establish the sBB as fastest on several small and medium-scale instances (not on the large ones, though). This is due to the guaranteed tight lower bound of zero which is known aprioristically. Knowledge of this MDGP-specific bound represents an exploitation of problem structure, and therefore gives sBB an advantage over the other methods.

2.2.2 Continuation methods

Following the ideas described in [39, 56], Moré and Wu proposed an algorithm, called DGSOL, based on a continuation approach for global optimization. The idea is to gradually transform the nonconvex, multimodal objective function (2) into a smoother function with fewer local minimizers, where an optimization algorithm is then applied to the transformed function, tracing their minimizers back to the original function. For other works based on continuation approach, see [7, 8, 21, 22, 23, 43].

The transformed function $\langle f\rangle_\lambda$, called the Gaussian transform, of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined by:

$$\langle f\rangle_\lambda(x) = \frac{1}{\pi^{n/2}\lambda^n}\int_{\mathbb{R}^n} f(y)\exp\left(-\frac{||y-x||^2}{\lambda^2}\right)dy, \tag{6}$$

where the parameter λ controls the degree of smoothing. The value $\langle f\rangle_\lambda(x)$ is a weighted average of $f(x)$ in a neighborhood of x, where the size of the neighborhood decreases as λ decreases: as λ tends to 0, the average is carried out on the singleton set $\{x\}$, thus recovering the original function in the limit. Smoother functions are obtained as λ increases.

This approach to the MDGP has been implemented and tested on two artificial instance types, where the molecule has $n = |V| = s^3$ atoms located in the three-dimensional lattice

$$\{(i_1, i_2, i_3) : 0 \le i_1 < s, 0 \le i_2 < s, 0 \le i_3 < s\}$$

for an integer $s \ge 1$. Computational results were obtained for $n \in \{27, 64, 125, 216\}$. In the first type, the atomic ordering is specified by letting i be the atom at the position (i_1, i_2, i_3), where $i = 1+i_1+si_2+s^2i_3$, and

$$E = \{\{i, j\} : |i-j| \le r\}, \tag{7}$$

where $r = s^2$. In the second model,

$$E = \{\{i, j\} : ||x_i - x_j|| \le \sqrt{r}\}, \tag{8}$$

where $x_i = (i_1, i_2, i_3)$ and $r = s^2$. For both models, s is considered in the interval $3 \le s \le 6$. In (7), E includes all nearby atoms, while in (8), E includes some of nearby atoms and some relatively distant atoms.

Each smoothed function in the sequence is solved using a globally convergent local NLP solution method, which is obviously a desirable feature. Continuation approaches seem to determine a global solution with less computational effort than is required by multistart type approaches.

2.2.3 Double VNS with smoothing

In [34], two efficient methods for large-scale molecules are combined into a heuristic method named *Double VNS with Smoothing* (DVS). It consists of two stages to be repeated until a certain termination condition becomes true. First, a smoothed version of the objective function (2) is derived as per [39] (see Sect. 2.2.2) and globally solved using VNS (see Sect. 2.2.1). The global optimum of the smoothed function is likely to be near the global optimum of the original problem, so a new VNS is deployed on (2) with tightened variable bounds. The computational results, comparing to DGSOL's performance, are good as regards accuracy (i.e. the value of the objective function (2) at the optimum), although DGSOL is much faster than DVS.

2.2.4 DC optimization

In [1, 2], An and Tao propose an approach for solving the exact MDGP, based on a d.c. (difference of convex functions) optimization algorithm. They work in $\mathcal{M}_{m,3}(\mathbb{R})$, the space of real matrices of order $n \times 3$, where $n = |V|$ and for each $Y \in \mathcal{M}_{n,3}(\mathbb{R})$, Y_i (respectively Y^i) is its i-th row (respectively i-th column). An embedding x is identified with the matrix Y by setting $Y_i^T = x_i$ for all $i \in V$. The MDGP can then be formulated as:

$$\min \left\{ \sigma(Y) = \frac{1}{2} \sum_{(i,j)\in S, i<j} w_{ij}\theta_{ij}(Y) : Y \in \mathcal{M}_{n,3}(\mathbb{R}) \right\}, \tag{9}$$

for some given weights $w_{ij} > 0$ for $i \neq j$ and $w_{ii} = 0$ for all $i \in V$. The pairwise potential $\theta_{ij} : \mathcal{M}_{n,3}(\mathbb{R}) \to \mathbb{R}$ is defined for problem (1) by either:

$$\theta_{ij}(Y) = \left(d_{ij}^2 - ||Y_i^T - Y_j^T||^2\right)^2 \tag{10}$$

or

$$\theta_{ij}(Y) = \left(d_{ij} - ||Y_i^T - Y_j^T||\right)^2, \tag{11}$$

and for problem (3) by θ_{ij} given by the $\{i,j\}$-th term of the objective function of (5). Y is a solution if and only if it is a global minimizer of problem (9) and $\sigma(Y) = 0$. We remark that although (9) may be nondifferentiable depending on the choice of θ_{ij}, it is also d.c.

An and Tao show that some d.c. algorithms can be adapted for developing efficient algorithms for solving large-scale exact MDGPs. They propose various versions of d.c. algorithms adapted to the different problem formulations. Although global optimality cannot be guaranteed for a general d.c. problem, the fact that global optimality can usually be obtained with a suitable starting point motivated the investigation of a technique for computing good starting points. These algorithms have been tested on three sets of data: the artificial data from Moré and Wu [39] (with up to 4096 atoms), 16 proteins in the PDB [5] (from 146 up to 4189 atoms), and the data from Hendrickson [19] (from 63 up to 777 atoms).

2.2.5 Alternating projections algorithm

The program APA, described in [44], puts together a sequence of existing algorithms. The main steps are as follows. First, a dissimilarity matrix $\Delta = (\delta_{ij})$ is generated randomly so that: (a) $\delta_{ij} = 0$ for all $\{i,j\} \notin E$ and $d_{ij}^L \leq \delta_{ij} \leq d_{ij}^U$ otherwise, and (b) Δ satisfies the triangular inequality. Secondly, Δ is projected onto the cone of matrices that are negative semidefinite on the orthogonal complement of the all-one vector, and projected back onto the data box, simultaneously zeroing the diagonal entries, obtaining Δ'. It can be shown that iterating this process an infinite number of times would yield a distance matrix Δ' satisfying (3). In practice, an embedding minimizing the distance from Δ' is obtained from the eigenvalues of Δ' after five iterations.

More precisely, APA rests on the following basic idea [15]. A symmetric $(n+1) \times (n+1)$ matrix $D = (d_{ij})$ with a 0 diagonal is a *Euclidean distance matrix* if there is $K \leq n$ and a vector function $x : \{0, \ldots, n\} \to \mathbb{R}^K$ such that $\forall 0 \leq i < j \leq n$ $\|x_i - x_j\| = d_{ij}$. By [49], this is equivalent to requiring that the $n \times n$ matrix $A = (a_{ij})$ defined by $a_{ij} = \frac{1}{2}(d_{0i} + d_{0j} - d_{ij})$ is positive semidefinite ($A \succeq 0$), and rk(A) is the minimum value of K ensuring the Euclidean distance matrix property for D. Furthermore, considering the spectral decomposition $A = U\Lambda U^\top$, where Λ is a diagonal matrix with the eigenvalues of A along the diagonal and letting $X = U\Lambda^{\frac{1}{2}}$, we have $A = XX^\top$ and the (i,k)-th component of X is x_{ik}, the k-th component of the K-vector x_i, for $i \in \{0,\ldots,n\}$. It is shown that $2A = P(-D)P^\top$ where P is $I - \frac{1}{n}\mathbf{1}(\mathbf{1}^\top)$ is the orthogonal projection on the subspace M orthogonal to $\mathbf{1}$, so D is a Euclidean distance matrix if and only if D is negative semidefinite on M. In APA, a projection operator having an equivalent effect as P is applied to the dissimilarity matrix Δ.

The computational results reported in [44] are all obtained from the bovine pancreatic trypsin inhibitor protein. The protein `1qlq` has 588 atoms including side chains. Accuracy-wise, APA is reported to yield Root Mean Square Deviation (RMSD) measures (given by $\|X - X'Q\|/\sqrt{K}$, where X, X' are two embedding matrices, whose i-th colums are $x_i, x'_i \in \mathbb{R}^K$ respectively, and Q is an appropriate rotation matrix) ranging in $O(1) - O(10)$.

2.2.6 Geometric build-up algorithm

In [11], Dong and Wu propose the solution of the exact MDGP by an algorithm, called the geometric build-up algorithm, based on a geometric relationship between coordinates and distances associated to the atoms of a molecule. It is assumed that it is possible to determine the coordinates of at least four atoms, which are marked as fixed; the remaining ones are non-fixed. The coordinates of a non-fixed atom a can be calculated by using the coordinates of four non-coplanar fixed atoms such that the distances between any of these four atoms and the atom a are known. If such four atoms are found, the atom a changes its status to fixed. More specifically, let b_1, b_2, b_3, b_4 be the four fixed atoms whose Cartesian coordinates are already known. Now suppose that the Euclidean distances among the atom a and the atoms b_1, b_2, b_3, b_4, namely d_{a,b_i}, for $i \in \{1,2,3,4\}$, are known. That is,

$$\begin{aligned} ||a - b_1|| &= d_{a,b_1}, \\ ||a - b_2|| &= d_{a,b_2}, \\ ||a - b_3|| &= d_{a,b_3}, \\ ||a - b_4|| &= d_{a,b_4}. \end{aligned}$$

Squaring both sides of these equations, we have:

$$\begin{aligned} ||a||^2 - 2a^T b_1 + ||b_1||^2 &= d^2_{a,b_1}, \\ ||a||^2 - 2a^T b_2 + ||b_2||^2 &= d^2_{a,b_2}, \\ ||a||^2 - 2a^T b_3 + ||b_3||^2 &= d^2_{a,b_3}, \\ ||a||^2 - 2a^T b_4 + ||b_4||^2 &= d^2_{a,b_4}. \end{aligned}$$

By subtracting one of these equations from the others, one obtaines a linear system that can be used to determine the coordinates of the atom a. For example, subtracting the first equation from the others, we obtain

$$Ax = b, \tag{12}$$

where

$$A = -2 \begin{bmatrix} (b_1 - b_2)^T \\ (b_1 - b_3)^T \\ (b_1 - b_4)^T \end{bmatrix},$$

$$x = a,$$

and

$$b = \begin{bmatrix} \left(d^2_{a,b_1} - d^2_{a,b_2}\right) - \left(||b_1||^2 - ||b_2||^2\right) \\ \left(d^2_{a,b_1} - d^2_{a,b_3}\right) - \left(||b_1||^2 - ||b_3||^2\right) \\ \left(d^2_{a,b_1} - d^2_{a,b_4}\right) - \left(||b_1||^2 - ||b_4||^2\right) \end{bmatrix}.$$

Since b_1, b_2, b_3, b_4 are non-coplanar atoms, the system (12) has a unique solution. If the exact distances between all pairs of atoms are given, this approach can determine the coordinates of all atoms of the molecule in linear time [10]. A further extension of the geometric build-up algorithm, dealing with the case of fewer than four known distances incident to any particular atom by means of a branching-type approach similar to that of the BP algorithm (Sect. 3.2.3), is given in [12]. Another development in the same direction is given in [55].

Dong and Wu report that the geometric build-up algorithm is very sensitive to the numerical errors introduced in computing the atomic coordinates. In [54], Wu and Wu propose an updated geometric build-up algorithm where the accumulated errors can be controlled. The latest implementation of this algorithm [54] was tested on a set of problems generated using the known structures of 10 proteins downloaded from the PDB [5], with problems from 404 up to 4201 atoms, yielding RMSD measures ranging from $O(10^{-8})$ to $O(10^{-13})$.

2.2.7 The GNOMAD iterative method

The GNOMAD algorithm [53] is an iterative method, based on a specific atomic order which changes iterationwise as each atom's contribution to the total error is updated. The method also exploits several local NLP searches (in low dimension) at each iteration. Particular attention is paid to the physically inviolable separation distances between atoms, that are usually referred to as Van der Waals distances. In fact, depending on the kind of atoms that interact, there is a minimum distance under which a repulsive force tends to separate such atoms. Therefore, if the aim is to find the stable conformation of a molecule, then it is preferable that all the Van der Waals distances are satisfied. This feature of the GNOMAD algorithm can also be exploited in other methods for the MDGP, even based on other approaches.

2.2.8 Monotonic Basin Hopping

In order to solve (5), a Monotonic Basin Hopping (MDH) algorithm is employed in [16]. The two key concepts in this algorithm are those of *funnel* and *funnel bottom*. Given a neighbourhood structure $\mathcal{N}$ of $\mathbb{R}^3$, a funnel is a maximal set $\mathcal{F}$ of local minima of the objective function (5) (call it $h(x)$) such that there exists a partial order $\sqsupset$ on $\mathcal{F}$ with the following property: for each $x \in \mathcal{F}$ there exists a finite descending chain $x = x_0 \sqsupset x_1 \sqsupset \ldots \sqsupset x_t = \min \mathcal{F}$ such that $h(x_j) > h(x_{j+1})$ and $x_{j+1} \in \mathcal{N}(x_j)$ for all $j < t$; x_t is called the *funnel bottom*. The exploration of different funnels can be performed with the aim of increasing the probability of catching a funnel whose bottom is the global optimum of (5). In [16], the authors propose a Population Basin Hopping (PDH) algorithm for the MDGP, in which several funnels are explored in parallel.

2.2.9 Semidefinite programming

The method described in [6] first forms vertex clusters that cover V in such a way that neighbouring clustering share some vertices (these are used to "stitch together" the embeddings restricted to each cluster). The clustering technique is based on permuting columns of the distance matrix (d_{ij}) so as to try to pool the nonzeros along the main diagonal. The partial embeddings for each cluster are computed by first solving an SDP relaxation of the quadratic system (3) restricted to edges in the cluster, and then applying a local NLP optimization algorithm that uses the optimal SDP solution as a starting point. When the distances have errors, there may not exist any valid embedding satisfying all the distance constraints. In this case, it is likely that the SDP approach (which relaxes these constraints anyhow) will end up yielding an embedding x' which is valid in a higher dimensional space $\mathbb{R}^{K'}$ where $K' > K$. In such cases, x' is projected onto an embedding x in $\mathbb{R}^K$. Such projected embeddings usually exhibit clusters of close vertices (none of which satisfies the corresponding distance constraints), due to correct distances in the higher dimensional space being "squeezed" to their orthogonal projection into the lower dimensional space. In order to counter this type of behaviour, a regularization objective $\max \sum_{i,j \in V} ||x_i - x_j||^2$ is added to the feasibility SDP.

2.2.10 A self-organization heuristic

The basic idea of the Stochasting Proximity Embedding (SPE) [57] heuristic is as follows. All the atoms are initially placed randomly into a cube of a given size. Pairs of atoms in E are repeatedly and randomly selected; for each pair $\{i, j\}$, the algorithm checks satisfaction of the corresponding constraint in (3). If the constraint is violated, the positions of the two atoms are changed according to explicit formulae in order to improve the current embedding. Naturally, since the algorithm works locally on pairs of atoms, there is no guarantee of obtaining a final solution satisfying all the constraints. Success stories concerning the SPE algorithm are reported in [20].

3 Search in discrete space

In this section we discuss discrete formulations and methods for the MDGP. We first present the topics which have had an influence on the discretization of the MDGP, then discuss a discretizable problem variant and the algorithmic framework used to solve it.

Notationwise, given a graph $G = (V, E)$, for all $v \in V$ we let $\delta(v) = \{u \in V \mid \{u, v\} \in E\}$. For a subset $U \subseteq V$ we let $G[U]$ be the subgraph of G induced by U, having edge set $E[U]$. For a totally ordered set $(V, <)$, for all $v \in V$ we let $\gamma(v) = \{u \in V \mid u < v\}$, and we define the *rank* of v as $\rho(v) = |\gamma(v)| + 1$.

3.1 The influence of rigidity

In the context of graph rigidity, embeddings are also called *realizations* [19]. A realization is *generic* if all vertex coordinates are algebraically independent over $\mathbb{Q}$. Although generic realizations are dense in the set of all realizations, as noted in [18], we really only need to avoid certain specific algebraic dependencies, so the genericity condition is not too hard to meet in practice. Given an undirected, weighted graph $G = (V, E, d)$ and a realization $x \in X$, the pair (G, x) is a *satisfying framework*. A *finite flexing* of a framework is an uncountable family $y \subseteq X$, indexed by a continuous parameter $t \geq 0$, such that there is t_0 with $x = y(t_0)$. Since for all t we have $y(t) \in X$, $(G, y(t))$ are all satisfying frameworks, thus for each edge $\{i, j\} \in E$, $||y_i(t) - y_j(t)||^2$ is constant. If there are no such families, the framework is *rigid*, otherwise it is *flexible*. We can differentiate the condition with respect to t [45] to obtain:

$$\forall \{i, j\} \in E \quad (v_i - v_j) \cdot (y_i - y_j) = 0, \tag{13}$$

where $v_i = \nabla_t y_i$ for all i. A function $v : E \to \mathbb{R}^3$ satisfying (13) is an *infinitesimal motion* of the framework. If there exists such a v then the framework is *infinitesimally flexible* and otherwise it is *infinitesimally rigid*. For generic realizations, infinitesimal rigidity implies rigidity. By a theorem of Gluck, if a graph has a single infinitesimally rigid realization, then all its generic realizations are infinitesimally rigid [14]. If we restrict attention to rigid realizations, by Gluck's theorem we can ignore the concept of framework and refer directly to *rigid graphs* [19]. The MDGP is still **NP**-hard even when restricted to rigid graphs [13].

3.1.1 ABBIE: a mixed discrete-continuous method

The concept of graph rigidity first made its way into the MDGP literature with the ABBIE [19] mixed discrete-continuous method for realizing general molecule graphs. Instead of solving (2) directly, ABBIE automatically finds the largest uniquely realizable rigid subgraphs of the molecule graphs and essentially contracts them to single vertices, yielding a graph minor G' with hopefully fewer vertices than the original graph. Sufficient conditions for unique graph realizability are given in [19]. The nonconvex MDGP problem (2) corresponding to G' is then solved using a multistart GO heuristic. The largest practical drawback of ABBIE is that, for interesting molecules such as proteins, the uniquely realizable subgraphs might fail to be large, yielding relatively insignificant CPU time reductions. The overall method fails to be exact because of the heuristic search in continuous space, but the combinatorial treatment using graph rigidity is a promising one, as we shall see in the remainder of the paper.

3.1.2 The Sensor Network Localization Problem

In [13], graph rigidity is used to investigate the SNLP:

> Sensor Network Localization Problem. Given an integer $K > 0$, a weighed undirected graph $G = (V, E, d)$ with $d : E \to \mathbb{R}_+$, a subset $U \subseteq V$, an embedding $x' : U \to \mathbb{R}^K$ s.t. $||x'_i - x'_j|| = d_{ij}$ for all $\{i, j\} \in E[U]$, find an extension $x : V \to \mathbb{R}^K$ of x' satisfying (1).

It is evident that SNLP$\supseteq$MDGP, for if $U = \emptyset$ then SNLP=MDGP: thus, a general method for solving the SNLP also solves the MDGP. The main algorithmic interest in [13] is to solve SNLP with $K = 2$, and to use the given x' to "grow" x iteratively exploiting graph rigidity by means of trilateration. The iteration follows a *trilateration order* of the vertices, i.e. an order $<$ on V such that:

1. letting U_0 be the set of the first $K + 1$ vertices in the order, $G[U_0]$ is the complete graph on $K + 1$ vertices;
2. for all $j > K + 1$, E contains at least $K + 1$ edges linking the j-th vertex to preceding vertices in the order.

If E is dense enough to grant the existence of a trilateration order, then G is rigid and x can be found in polynomial time [13].

Although the above method seems to be extremely interesting in network analysis, no-one has yet found an interesting class of molecules for which a trilateration order on the atoms can be established *a priori*.

3.2 A discrete MDGP variant

Borrowing ideas from rigidity and SNLP, [32, 33] define a subproblem of the MDGP which includes proteins and for which the search is completely discrete. Given a graph $G = (V, E)$ with an order $<$ on the vertex set V, let $U_0 = \{v \in V \mid \rho(v) \le K\}$ be the set of the first K vertices of $(V, <)$ and for all $v \in V$ having rank $\rho(v) > K$ let $U_v = \{u \in V \mid \rho(v) - K \le \rho(u) \le \rho(v)\}$ be the subset of vertices including v and its K immediate predecessors.

> Discretizable Molecular Distance Geometry Problem (DMDGP). Given a weighted undirected graph $G = (V, E, d)$ where $d : E \to \mathbb{R}_+$ and a total order $<$ on V such that:
>
> 1. $G[U_0]$ is the clique on K vertices (starting configuration);
> 2. for all v s.t. $\rho(v) > K$, $G[U_v]$ is the clique on $K + 1$ vertices (discretizing order);
> 3. for all v s.t. $\rho(v) > K$ and all subsets $\{u, w, z\}$ with ranks in $\rho(v) - K, \ldots, \rho(v) - 1$ we have $d_{uz} < d_{uw} + d_{wz}$ (strict triangular inequality),
>
> find an embedding $x : V \to \mathbb{R}^K$ such that (1) holds.

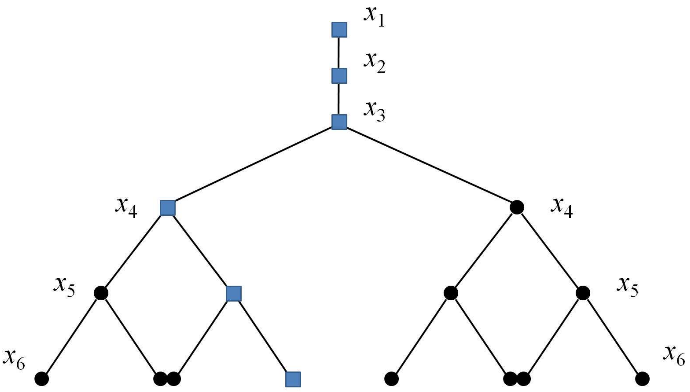

Figure 1: The binary tree T for $n = 6$. The boxes show a complete path on T, which corresponds to a possible solution to the DMDGP.

The order $<$ given as part of the data in the DMDGP can be successfully used to construct a (partial) directed binary tree T of depth $n = |V|$ each node of which, at level $i \leq n$, represents a possible spatial position x_i for the i-th atom. Thus, each path of length n in T represents an embedding of G in $\mathbb{R}^K$ (see Figure 1).

Notice that the discretizing order falls short of the trilateration order defined in Sect. 3.1.2 by exactly one vertex: requiring U_v to contain $K + 2$ vertices and $G[U_v]$ to be the complete graph on $K + 2$ vertices would be a sufficient condition for the existence of a trilateration order, in the presence of which the MDGP is known to be polynomial. The binary tree construction can occur because, at a generic atom $v \in V$ with $\rho(v) = i > K$, the vector d includes distances to v from atoms $i - K, \ldots, i - 1$. Assuming all atoms in $\gamma(v)$ have already been positioned, v belongs to the intersection of K spheres in $\mathbb{R}^K$. Because of strict triangular inequality, this intersection consists of at most two points.

In order to see this in the case $K = 3$, recall that the intersection of two spheres (spherical surfaces) in $\mathbb{R}^3$ is either empty, or a single point, or a circle in space; intersecting these sets with a third sphere might yield the empty set, a singleton, two distinct points, or the whole circle. The latter case, however, corresponds to a configuration of 3 collinear atoms, which is impossible by the strict triangular inequality (see Fig. 2). This idea is at the core of the main solution technique used in the solution of the DMDGP — the Branch-and-Prune algorithmic framework. We remark that the trilateration order given by [13] yields a polynomial algorithm simply because adding a vertex to U_v would make v stand at the intersection of *four* spheres in $\mathbb{R}^3$, which in presence of the strict triangular inequality is either empty or a singleton, thereby reducing T to a single path.

Algebraically, if we know the position of three atoms $x_1, x_2, x_3 \in \mathbb{R}^3$ and the distances d_1, d_2, d_3 to a fourth atom y, we obtain the system:

$$\begin{aligned} ||x_1 - y||^2 &= d_1^2 \\ ||x_2 - y||^2 &= d_2^2 \\ ||x_3 - y||^2 &= d_3^2. \end{aligned}$$

We subtract the third equation from the first and the second, to get:

$$\begin{aligned} 2(x1 - x3) \cdot y &= (||x_1||^2 - d_1^2) - (||x_3||^2 - d_3^2) && (14) \\ 2(x2 - x3) \cdot y &= (||x_2||^2 - d_1^2) - (||x_3||^2 - d_3^2) && (15) \\ ||x_3 - y||^2 &= d_3^2. && (16) \end{aligned}$$

The linear part (14)-(15) can be written as $Ay = b$ where A is a 2×3 matrix. Let B, N be a partition of the columns of A into basics and nonbasics; we can then write $y_B = (A_B)^{-1}(b - A_N y_N)$. If A has full

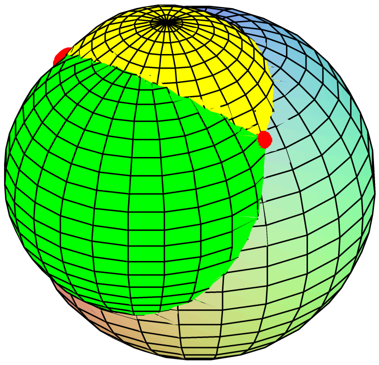

Figure 2: The intersection of three spherical surfaces in $\mathbb{R}^3$, containing exactly two points.

rank then $y_N \in \mathbb{R}^1$, hence by replacing y_B in (16) we obtain a single quadratic equation in one variable, which has at most two solutions. If $\mathrm{rk}(A) < 2$, either $\mathrm{rk}(A) = 0$ (which means that $x_1 = x_2 = x_3$, which in particular implies that x_1, x_2, x_3 are collinear, violating the strict triangular inequality) or $\mathrm{rk}(A) = 1$, which implies $x_1 - x_3$ and $x_2 - x_3$ are linearly dependent, i.e. x_1, x_2, x_3 are collinear, again violating strict triangular inequality.

3.2.1 Relevance to proteins

Whereas the polynomial method in [13] requires conditions that are too strict to be applied in practice, it is very likely that most protein backbones satisfy the DMDGP definition. Proteins are molecules whose atoms can be partitioned into two distinct sets: the *backbone* atoms and the *side chains* atoms. The problem of positioning side chains given the backbone embedding is known as the Side Chain Placement Problem (SCPP) [46, 47] and is **NP**-hard.

The protein backbone induces a natural total order on V. Since quadruplets of subsequent atoms are spatially close, this order looks like a good candidate for the order required by the DMDGP definition. The distance from a generic atom $v \in V$ with rank $\rho(v) > 3$ to its preceding atom in the backbone can be known because of chemical reason, and so is the angle defined by the three atoms ranked $\rho(v)-2, \rho(v)-1, \rho(v)$. Since the distance between the atoms ranked $\rho(v)-2, \rho(v)-1$ can also be known because of chemical reasons, the distance between the atoms ranked $\rho(v)-2$ and $\rho(v)$ could be computed. Furthermore, since the distance between atoms ranked $\rho(v)-3$ and $\rho(v)$ is usually smaller than 6Å, it is very likely to be estimated by NMR experiments (the NMR distance threshold is usually assumed to be somewhere between 5 and 6Å). Thus, DMDGP conditions 1-2 are satisfied in most of the cases. As for 3, no protein exhibiting a backbone with three exactly collinear atoms has ever been found yet.

3.2.2 Complexity of DMDGP

It was proved in [32] that the DMDGP is **NP**-hard with a reduction from Subset-Sum. The proof is similar in spirit to that provided in [48] for the 1-dimensional case of the MDGP.

3.2.3 The Branch-and-Prune algorithmic framework

The BP algorithm explores the tree T mentioned at the beginning of Sect. 3.2 (p. 10), branching the search whenever the next atom in the order can take exactly two positions. We remark that the DMDGP definition involves distances from a generic atom $v \in V$ to K preceding atoms in the order; protein backbones, however, due to twisting and coiling, often display spatial proximity between atoms whose ranks in the order are very different. Since these small distances will very likely also be estimated by NMR experiments, the corresponding edges are expected to be in E; these distances can be used in BP to prune large parts of the search tree. The largest possible binary tree search for an embedding in X will have 2^{i-K} nodes at level $i \leq n = |V|$. A known distance $\{j, i\}$ where $j < i - K$, however, may be infeasible with respect to many of the 2^{i-K} worst-case partial embeddings, and could potentially make the search tree collapse to a few possible nodes at level i.

We now restrict our attention to $K = 3$ for simplicity, and without (excessive) loss of generality. Algorithm 1 provides a sketch of the BP algorithm for the exact DMDGP. BP is a recursive procedure deployed at each node of the tree T. Its arguments are as follows:

- i is the rank of the atom whose position is to be determined;
- $x_{<i}$ is the partial embedding represented by the path on the tree T from the root to the current node;
- $G = (V, E, d)$ is an encoding of the instance;
- $\hat{X}$ is a collection of valid embeddings for G found by BP.

Notationwise, we let $S(y, r)$ be the sphere in $\mathbb{R}^3$ centered in $y \in \mathbb{R}^3$ with radius $r \in \mathbb{R}$.

Algorithm 1 The BP algorithmic framework.

```
1: BranchAndPrune($i$, $x_{<i}$, $G$, $\hat{X}$)
2: Let $\mathcal{S} \leftarrow \bigcap_{k \leq 3} S(x_{i-k}, d_{i-k,i}) = \{s_1, \ldots, s_q\}$, where $q \in \{1, 2\}$
3: for $p \leq q$ do
4:    Extend the current embedding to $x^{(p)} = (x_{<i}, s_p)$
5:    if $x^{(p)}$ satisfies (1) then
6:       if $(i = n)$ then
7:          Let $\hat{X} \leftarrow \hat{X} \cup \{x^{(p)}\}$
8:       else
9:          BranchAndPrune($i + 1$, $x^{(p)}$, $G$, $\hat{X}$)
10:      end if
11:   end if
12: end for
```

The algorithm is initially invoked with $i = 4$, $x_{<4}$ being an initial partial embedding for the first three atoms, and $\hat{X} = \emptyset$. It calls itself recursively on increasing values of i, building up embeddings as it dynamically generates the nodes of T. Notice that whenever the intersection of three spheres is given by two points, two subnodes of the current nodes are generated.

We call Alg. 1 an algorithmic framework because it does not specify how to check whether $x^{(p)}$ satisfies (1) on Line 5. We already remarked earlier that $\mathcal{S}$, defined on Line 2, has at most two points. Theorem 3.1 is shown to hold in [33].

3.1 Theorem ([33])
At termination of Alg. 1, $\hat{X} = X$.

There are different ways for checking the feasibility of computed atomic positions at Line 5 of Alg. 1. If exact distances are available, then the most natural pruning test is the one in which, every time a new position x_i is computed for atom i, the subset of constraints in (1) involving the index i are checked. The position is feasible only if these constraints are satisfied. Since equations cannot be verified exactly in floating point arithmetic, it is important to set up a tolerance ε accurately, because excessively small tolerances could force the pruning of all the atomic positions, whereas excessively large ones could allow infeasible positions to be accepted. Thus, it is unfortunately the case that Thm. 3.1 only holds in an ideal case where equations in (1) can be checked exactly.

A different pruning test is based on point-to-point shortest paths in G. Consider atoms h, i, k with $h < i < k$ such that $(h, k) \in E$, i.e. the distance d_{hk} is known. Let us suppose that the BP algorithm already placed atom h, and the feasibility of the atom i needs to be verified. Let $D(i, k)$ be an upper bound to the distance $||x_i - x_k||$ for all possible solutions to the problem. It has been proved in [32] that, if the converse triangular inequality $D(i, k) + d_{hk} < ||x_h - x_i||$ holds, then the current BP node establishing the position of atom i can be pruned. One way for computing a reasonably tight upper bound $D(i, k)$ to $||x_i - x_k||$ is by finding the shortest path between vertex i and vertex k of the graph G.

When both pruning tests are used together [28], the BP algorithm is able to find valid embeddings in a smaller number of steps. This apparent efficiency, however, is not reflected on the computational cost needed to apply it. While the natural pruning test just checks that distances are contained in certain intervals in $O(K)$, essentially $O(1)$ if $K = 3$, shortest path computations are of order $O(n^2)$ in general. Even precomputing all shortest paths ($O(n^3)$ but carried out only once) does not seem to reverse this trend in practice.

Computational experiments [28, 29, 32, 40, 41, 42] showed that the BP algorithm is able to efficiently solve instances of the DMDGP and that it compares well with other algorithms based on a continuous formulation of the problem, both w.r.t. computational efficiency and accuracy. We remark that the BP algorithm can be easily extended to deal with the inexact version of the DMDGP, i.e. when exact distances are replaced by bounds.

3.3 Future challenges

Distance geometry has been for decades the mathematical tool of choice when working out tridimensional molecular structure from NMR data. Methods searching the (continuous) Euclidean space are useful when no further molecular structure is known. For the all-important class of proteins, however, the search can be discretized, and tools such as the BP algorithmic framework clearly outperform continuous approaches on both accuracy and CPU time. We recall that the embedding problem for protein graphs in $\mathbb{R}^3$ consists in two stages: placing the backbone and then placing the side chains; and that independent research is conducted on both fronts. The DMDGP conditions are verified for most protein backbones, and so in principle the technique should now be applicable to real proteins. There is, however, a last difficulty to overcome: NMR data can usually measure inter-atomic distance *between hydrogen atoms*, whereas protein backbones do *not* only consists of hydrogen atoms. The current effort, conducted by the authors of this survey, is to identify a virtual hydrogen backbone on the whole protein on which the DMDGP order exists. As long as the backbone is not given but must be identified, there is also an opportunity to relax the DMDGP definition somewhat, and to only require that the order be such that each atom has at least three adjacent precedents, which need not be immediate precedessors (similarly to the trilateration order). This opens up new combinatorial problems: given the graph G, identify an order satisfying the DMDGP definition, or a relaxed order as described above.

Another challenge, of a purely mathematical nature, is to explain why computational experiments on the DMDGP always return a set of embeddings $\hat{X}$ with cardinality a power of two. Work based on the study of DMDGP symmetries is currently ongoing in this direction.

4 Conclusion

The purpose of this survey is to show how research on the Molecular Distance Geometry Problem, which establishes embeddings of molecules in Euclidean space using NMR data, evolved from purely continuous search methods to mixed-discrete (via concepts in graph rigidity) to almost exclusively combinatorial — which can be applied to proteins. Although there are more continuous methods in the literature than discrete ones, many of the continuous approaches involve a combinatorial element at some degree.

It is an interesting fact that MDGP solution methods started off in chemical communities and later moved to the field of mathematics (geometry and optimization). Nowadays most MDGP papers, including those written by the authors of this survey, report computational experiments conducted on publically available proteins from the PDB where the NMR experiments are simulated: only those distances smaller than 6Å are kept. It is these authors' opinion that current discrete DMDGP techniques are almost up to the challenge of tackling *real* NMR protein data, thereby moving back to chemistry and biochemistry.

Acknowledgments

The authors would like to thank for financial support: the Brazilian research agencies FAPESP and CNPq. One of the authors (CL) is grateful to the French research agency CNRS and École Polytechnique.

References

[1] L.T. Hoai An. Solving large scale molecular distance geometry problems by a smoothing technique via the gaussian transform and d.c. programming. *Journal of Global Optimization*, 27:375–397, 2003.

[2] L.T. Hoai An and P.D. Tao. Large-scale molecular optimization from distance matrices by a d.c. optimization approach. *SIAM Journal on Optimization*, 14:77–114, 2003.

[3] J. Bachrach and C. Taylor. Localization in sensor networks. In I. Stojmenović, editor, *Handbook of Sensor Networks*. Wiley, 2005.

[4] P. Belotti, J. Lee, L. Liberti, F. Margot, and A. Wächter. Branching and bounds tightening techniques for non-convex MINLP. *Optimization Methods and Software*, 24(4):597–634, 2009.

[5] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acid Research*, 28:235–242, 2000.

[6] P. Biswas, K.C. Toh, and Y. Ye. A distributed SDP approach for large-scale noisy anchor-free graph realization with applications to molecular conformation. *SIAM Journal on Scientific Computing*, 30(3):1251–1277, 2008.

[7] T.F. Coleman, D. Shalloway, and Z. Wu. Isotropic effective energy simulated annealing searches for low energy molecular cluster states. *Computational Optimization and Applications*, 2:145–170, 1993.

[8] T.F. Coleman, D. Shalloway, and Z. Wu. A parallel build-up algorithm for global energy minimizations of molecular clusters using effective energy simulated annealing. *Journal of Global Optimization*, 4:171–185, 1994.

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

[10] Q. Dong and Z. Wu. A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances. *Journal of Global Optimization*, 22:365–375, 2002.

[11] 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.

[12] R. dos Santos Carvalho, C. Lavor, and F. Protti. Extending the geometric build-up algorithm for the molecular distance geometry problem. *Information Processing Letters*, 108:234–237, 2008.

[13] T. Eren, D.K. Goldenberg, W. Whiteley, Y.R. Yang, A.S. Morse, B.D.O. Anderson, and P.N. Belhumeur. Rigidity, computation, and randomization in network localization. *IEEE Infocom Proceedings*, pages 2673–2684, 2004.

[14] H. Gluck. Almost all simply connected closed surfaces are rigid. In *Geometric Topology*, volume 438 of *Lecture Notes in Mathematics*, pages 225–239, Berlin, 1975. Springer.

[15] W. Glunt, T.H. Hayden, S. Hong, and J. Wells. An alternating projection algorithm for computing the nearest euclidean distance matrix. *SIAM Journal on Matrix Analysis and Applications*, 11(4):589–600, 1990.

[16] A. Grosso, M. Locatelli, and F. Schoen. Solving molecular distance geometry problems by global optimization algorithms. *Computational Optimization and Applications*, 43:23–27, 2009.

[17] P. Hansen and N. Mladenović. Variable neighbourhood search: Principles and applications. *European Journal of Operations Research*, 130:449–467, 2001.

[18] B.A. Hendrickson. Conditions for unique graph realizations. *SIAM Journal on Computing*, 21(1):65–84, 1992.

[19] B.A. Hendrickson. The molecule problem: exploiting structure in global optimization. *SIAM Journal on Optimization*, 5:835–857, 1995.

[20] S. Izrailev, F. Zhu, and D.K. Agrafiotis. A distance geometry heuristic for expanding the range of geometries sampled during conformational search. *Journal of Computational Chemistry*, 26(3):1962–1969, 2006.

[21] J. Kostrowicki and L. Piela. Diffusion equation method of global minimization: performance for standard functions. *Journal of Optimization Theory and Applications*, 69:269–284, 1991.

[22] J. Kostrowicki, L. Piela, B.J. Cherayil, and H.A. Scheraga. Performance of the diffusion equation method in searches for optimum structures of clusters of lennard-jones atoms. *Journal of Physical Chemistry*, 95:4113–4119, 1991.

[23] J. Kostrowicki and H.A. Scheraga. Application of the diffusion equation method for global optimization of oligopeptides. *Journal of Physical Chemistry*, 96:7442–7449, 1992.

[24] S. Kucherenko and Yu. Sytsko. Application of deterministic low-discrepancy sequences in global optimization. *Computational Optimization and Applications*, 30(3):297–318, 2004.

[25] C. Lavor. On generating instances for the molecular distance geometry problem. In Liberti and Maculan [35], pages 405–414.

[26] C. Lavor, L. Liberti, and N. Maculan. Computational experience with the molecular distance geometry problem. In J. Pintér, editor, *Global Optimization: Scientific and Engineering Case Studies*, pages 213–225. Springer, Berlin, 2006.

[27] C. Lavor, L. Liberti, and N. Maculan. Molecular distance geometry problem. In C. Floudas and P. Pardalos, editors, *Encyclopedia of Optimization*, pages 2305–2311. Springer, New York, 2 edition, 2009.

[28] C. Lavor, L. Liberti, A. Mucherino, and N. Maculan. On a discretizable subclass of instances of the molecular distance geometry problem. In *Proceedings of the 24th Annual ACM Symposium on Applied Computing*, pages 804–805. ACM, 2009.

[29] C. Lavor, A. Mucherino, L. Liberti, and N. Maculan. Computing artificial backbones of hydrogen atoms in order to discover protein backbones. In *Proceedings of the International Multiconference on Computer Science and Information Technology, Workshop on Computational Optimization*. IEEE, 2009.

[30] L. Liberti. Writing global optimization software. In Liberti and Maculan [35], pages 211–262.

[31] L. Liberti and M. Dražic. Variable neighbourhood search for the global optimization of constrained NLPs. In *Proceedings of GO Workshop, Almeria, Spain*, 2005.

[32] L. Liberti, C. Lavor, and N. Maculan. A branch-and-prune algorithm for the molecular distance geometry problem. Technical Report q-bio/0608012, arXiv, 2006.

[33] 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.

[34] L. Liberti, C. Lavor, N. Maculan, and F. Marinelli. Double variable neighbourhood search with smoothing for the molecular distance geometry problem. *Journal of Global Optimization*, 43:207–218, 2009.

[35] L. Liberti and N. Maculan, editors. *Global Optimization: from Theory to Implementation*. Springer, Berlin, 2006.

[36] L. Liberti, N. Mladenović, and G. Nannicini. A good recipe for solving MINLPs. In V. Maniezzo, T. Stützle, and S. Voß, editors, *Hybridizing metaheuristics and mathematical programming*, volume 10 of *Annals of Information Systems*, New York, 2009. Springer.

[37] L. Liberti, P. Tsiakis, B. Keeping, and C.C. Pantelides. *oo$\mathcal{OPS}$*. Centre for Process Systems Engineering, Chemical Engineering Department, Imperial College, London, UK, 2001.

[38] N. Mladenović, J. Petrović, V. Kovačević-Vujčić, and M. Čangalović. Solving a spread-spectrum radar polyphase code design problem by tabu search and variable neighbourhood search. *European Journal of Operations Research*, 151:389–399, 2003.

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

[40] A. Mucherino and C. Lavor. The branch and prune algorithm for the molecular distance geometry problem with inexact distances. In *Proceedings of the International Conference on Computational Biology*, volume 58, pages 349–353. World Academy of Science, Engineering and Technology, 2009.

[41] A. Mucherino, C. Lavor, and N. Maculan. The molecular distance geometry problem applied to protein conformations. In S. Cafieri, A. Mucherino, G. Nannicini, F. Tarissan, and L. Liberti, editors, *Proceedings of the* 8^{th} *Cologne-Twente Workshop on Graphs and Combinatorial Optimization*, pages 337–340, Paris, 2009. École Polytechnique.

[42] A. Mucherino, L. Liberti, C. Lavor, and N. Maculan. Comparisons between an exact and a metaheuristic algorithm for the molecular distance geometry problem. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 333–340, Montreal, 2009. ACM.

[43] L. Piela, J. Kostrowicki, and H.A. Scheraga. The multiple-minima problem in the conformational analysis of molecules: deformation of the protein energy hypersurface by the diffusion equation method. *Journal of Physical Chemistry*, 93:3339–3346, 1989.

[44] R. Reams, G. Chatham, W. Glunt, D. McDonald, and T. Hayden. Determining protein structure using the distance geometry program APA. *Computers and Chemistry*, 23:153–163, 1999.

[45] B. Roth. Rigid and flexible frameworks. *American Mathematical Monthly*, 88(1):6–21, 1981.

[46] R. Santana, P. Larrañaga, and J.A. Lozano. Combining variable neighbourhood search and estimation of distribution algorithms in the protein side chain placement problem. In *Proc. of Mini Euro Conference on Variable Neighbourhood Search, Tenerife, Spain*, 2005.

[47] R. Santana, P. Larrañaga, and J.A. Lozano. Combining variable neighbourhood search and estimation of distribution algorithms in the protein side chain placement problem. *Journal of Heuristics*, 14:519–547, 2008.

[48] J.B. Saxe. Embeddability of weighted graphs in k-space is strongly NP-hard. *Proceedings of 17th Allerton Conference in Communications, Control and Computing*, pages 480–489, 1979.

[49] I.J. Schoenberg. Remarks to maurice fréchet's article "sur la définition axiomatique d'une classe d'espaces distanciés vectoriellement applicable sur l'espace de hilbert". *Annals of Mathematics*, 36(3):724–732, 1935.

[50] E.M.B. Smith and C.C. Pantelides. A symbolic reformulation/spatial branch-and-bound algorithm for the global optimisation of nonconvex MINLPs. *Computers & Chemical Engineering*, 23:457–478, 1999.

[51] M-C. So and Y. Ye. Theory of semidefinite programming for sensor network localization. *Mathematical Programming*, 109:367–384, 2007.

[52] M. Tawarmalani and N.V. Sahinidis. Global optimization of mixed integer nonlinear programs: A theoretical and computational study. *Mathematical Programming*, 99:563–591, 2004.

[53] G.A. Williams, J.M. Dugan, and R.P. Altman. Constrained global optimization for estimating molecular structure from atomic distances. *Journal of Computational Biology*, 8:523–547, 2001.

[54] D. Wu and Z. Wu. An updated geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data. *Journal of Global Optimization*, 37:661–673, 2007.

[55] D. Wu, Z. Wu, and Y. Yuan. Rigid versus unique determination of protein structures with geometric buildup. *Optimization Letters*, 2(3):319–331, 2008.

[56] Z. Wu. The effective energy transformation scheme as a special continuation approach to global optimization with application to molecular conformation. *SIAM Journal on Optimization*, 6:748–768, 1996.

[57] H. Xu, S. Izrailev, and D.K. Agrafiotis. Conformational sampling by self-organization. *Journal of Chemical Information and Computer Sciences*, 43:1186–1191, 2003.