Noname manuscript No.
(will be inserted by the editor)

On the Computation of Protein Backbones by using Artificial Backbones of Hydrogens

C. Lavor · A. Mucherino · L. Liberti · N. Maculan

Received: date / Accepted: date

Abstract NMR experiments provide information from which some of the distances between pairs of hydrogen atoms of a protein molecule can be estimated. Such distances can be exploited in order to identify the three-dimensional conformation of the molecule: this problem is known in the literature as the Molecular Distance Geometry Problem (MDGP). In this paper, we show how an artificial backbone of hydrogens can be defined which allows the reformulation of the MDGP as a combinatorial problem. This is done with the aim of solving the problem by the Branch and Prune (BP) algorithm, which is able to solve it efficiently. Moreover, we show how the real backbone of a protein conformation can be computed by using the coordinates of the hydrogens found by the BP algorithm. Formal proofs of the presented results are provided, as well as computational experiences on a set of instances whose size ranges from 60 to 6000 atoms.

Keywords distance geometry · protein molecules · hydrogen atoms · combinatorial optimization · branch and prune

1 Introduction

Proteins are important molecules because they perform different functions, often of vital importance, in the cells of the living beings. Their function is determined by the dynamics of the proteins, which depends on their three-dimensional conformation. While finding the chemical composition of a protein molecule is nowadays relatively

C. Lavor
Dept. of Applied Mathematics (IMECC-UNICAMP), State University of Campinas, Campinas-SP, Brazil, E-mail: clavor@ime.unicamp.br

A. Mucherino
INRIA Lille Nord Europe, Villeneuve d'Ascq, France, E-mail: antonio.mucherino@inria.fr

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

N. Maculan
COPPE, Systems Engineering, Federal University of Rio de Janeiro, Rio de Janeiro-RJ, Brazil, E-mail: maculan@cos.ufrj.br

simple, finding its three-dimensional conformation is not an easy task. The Nuclear Magnetic Resonance (NMR) is an experimental technique that is able to provide information from which some of the distances between pairs of atoms forming the molecule can be estimated [9]. These experimentally obtained data can then be used for computing the coordinates (into a given Cartesian system) of all the atoms of the molecule. The problem of finding the three-dimensional conformation of the molecule (i.e. the coordinates of all its atoms), starting from the known distances between pairs of atoms, is referred to as the Molecular Distance Geometry Problem (MDGP) [5]. The focus of this paper is, in particular, on protein molecules.

Over the years, many methods have been proposed for solving the MDGP. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a protein conformation, where $x_i \in \mathbb{R}^3$ represents the i^{th} atom of the protein, in a given ordering. Let E be the set of pairs of atoms whose distance is known. Then, the MDGP can be seen as the problem of finding X such that

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

where $||x_i - x_j||$ represents the computed distance between two atoms of X, and d_{ij} is the known value of the generic distance. The decision version of this problem (is there an X such that (1) holds?) is **NP**-hard by [28]. The MDGP is usually cast as a global optimization problem, where the aim is to minimize an objective function which is able to provide a measure of how much the distances $||x_i - x_j||$, related to a certain conformation X, differ from the known distances d_{ij}, for each $(i,j) \in E$. Different objective functions have been proposed, and one of the most often used is the Largest Distance Error (LDE):

$$LDE(X) = \frac{1}{m} \sum_{\{i,j\}} \frac{|\, ||x_i - x_j|| - d_{ij} \,|}{d_{ij}}, \tag{2}$$

where m is the total number of known distances. Supposing that a position is given to all the atoms of the conformation X, if the value of the LDE function is 0, then the set of given distances is feasible and the conformation X satisfies all of them.

Different methods have been proposed over time for solving this global optimization problem. One of the difficulties to be faced is that the LDE function (and even other penalty functions used in this context) has many local minima, where a method for optimization can get stuck at. In order to overcome this problem, Moré and Wu [21,22] proposed to approximate the used penalty function with smoother functions converging to the original one. Another method for the MDGP makes use of a penalty function which can be seen as the difference of two convex functions [1], and particular techniques for *d.c.* optimization are used. Moreover, in [8], a Population Basin Hopping method is employed, in which basic concepts (such as the ones of *funnel* and *funnel bottom*) usually used in methods for finding molecular conformations are exploited. Note that many of these methods are based on a continuous formulation of the problem, and that deterministic methods are often employed for their solution. However, there are heuristic algorithms particularly designed for solving the MDGP, such as, for example, the SPE algorithm [33] and the one proposed in [19]. Other algorithms, as for example [30], are instead based on a meta-heuristic search, such as Simulated Annealing [10]. For a survey on methods and algorithms for the MDGP, the reader is refereed to [11, 20].

Even though many algorithms have been proposed for solving MDGPs, there are actually only a few implementations of such algorithms that are freely available to

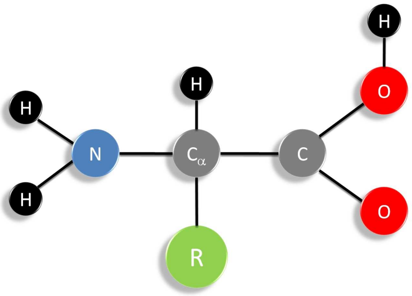

Fig. 1 The general structure of an amino acid.

the scientific community. For example, the software DGSOL implements the algorithm proposed by Moré and Wu [21, 22], and it is freely downloadable from the first author's web page. Other examples of free software for MDGP are Xplor-NIH [30] and TINKER (`http://dasher.wustl.edu/tinker/`). Both of them are based on meta-heuristic searches.

Recently, a new approach to the MDGP has been proposed. In the event that some particular assumptions are satisfied (see Section 2), the global optimization problem associated to the MDGP is reformulated as a combinatorial optimization problem. In this way, the search domain is reduced to a discrete set. Computational experiments presented, for example, in [12, 14–18, 25], showed that the combinatorial approach to the MDGP is more efficient than the continuous one. We refer to this combinatorial reformulation of the MDGP as the Discretizable Molecular Distance Geometry Problem (DMDGP). A software based on a Branch & Prune (BP) algorithm [11], which we employ for solving instances of the DMDGP, is currently under development [24]. `MD-jeep` is freely available to the scientific community, and we plan to integrate the software with the results presented in this paper in the future releases.

Proteins are chains of smaller molecules called *amino acids*, which are chemically bound to each other through a subgroup of atoms that each amino acid has in common. We will refer to this subgroup of atoms as the *common part* of each amino acid. All these parts define the so-called *backbone* of the protein. When two amino acids bind to each other during protein synthesis, some of the atoms of their common parts are lost, and the carbon atom C of the first amino acid binds to the nitrogen N of the second one. Therefore, the protein backbone is finally formed by the sequence of atoms $N - C_\alpha - C$, where oxygen and hydrogen atoms are also attached (see Figure 1).

In previous works [12, 14, 18, 25], the DMDGP has been tested on instances related to the sequence of atoms $N - C_\alpha - C$ of the protein backbones. As it is also supposed in many related works (see for example [3, 4, 31, 32]), no distinctions between the different kinds of atoms (N, C, H, O, ...) were considered. The computational experiments showed that instances related to the sequence $N - C_\alpha - C$ can be almost always solved by the combinatorial approach, because the necessary assumptions are satisfied.

However, the majority of the distances obtained by NMR are distances between pairs of hydrogens. In order to perform more realistic experiments, we do not consider in this work the sequence $N - C_\alpha - C$, but rather all the hydrogen atoms of the protein backbones. Unfortunately, in general, if only the hydrogens of the protein backbones are considered in the natural ordering given to these atoms, then the corresponding instances do not satisfy the assumptions of the DMDGP.

The focus of this paper is two-fold. First of all, we will show how to identify artificial backbones of hydrogen atoms such that the needed assumptions for formulating the problem as a DMDGP are satisfied. Preliminary studies on artificial backbones of hydrogens can be found in [15,16,27]. Secondly, we will show how the real backbone of a protein conformation (the sequence $\mathrm{N} - \mathrm{C}_\alpha - \mathrm{C}$) can be computed by exploiting the coordinates of the hydrogens obtained by solving the DMDGP related to a given artificial backbone of hydrogens. In the computational experiments, we will work in the simplified case in which the distances can be considered as accurate. However, the work here presented can be extended in order to manage experimental errors, by integrating, for example, the strategies presented in [23,26].

The paper is organized as follows. In Section 2, we will give an outline of the DMDGP and of the BP algorithm. Emphasis will be given to the assumptions that must be satisfied in order to formulate the problem as a DMDGP. In Section 3, we will show how to generate an artificial backbone of hydrogens that satisfies the necessary assumptions. In Section 4, we will show how to generate the real backbone $\mathrm{N} - \mathrm{C}_\alpha - \mathrm{C}$ of a protein conformation by using the coordinates of the hydrogens of an artificial backbone, once they have been obtained by applying the BP algorithm. Finally, computational experiences are shown in Section 5, and conclusions are drawn in Section 6.

2 The DMDGP and the Branch and Prune algorithm

Let us suppose that some of the distances between pairs of atoms of a molecule are known. Let $G = (V, E, d)$ be a weighted undirected graph, where

- there is a vertex $i \in V$ associated to each atom of the molecule;
- there is an edge $(i, j) \in E$ if and only if the distance between i and j is known;
- the weights d associated to the edges provide the numerical values of the known distances.

The Molecular Distance Geometry Problem (MDGP) is the problem of finding an embedding $x : V \to \mathbb{R}^3$ such that the molecular conformation

$$X = \{x(i) : i \in V\}$$

satisfies all the distances d.

The MDGP can be formulated as a combinatorial problem if there exists an ordering on V such that the following two assumptions are satisfied:

Assumption 1: all the distances $d_{i-3,i}$, $d_{i-2,i}$ and $d_{i-1,i}$ are known, for each i;
Assumption 2: for each triplet of vertices $\{i-2, i-1, i\}$, the strict triangular inequality

$$d_{i-2,i} < d_{i-2,i-1} + d_{i-1,i}$$

holds,

where the subtraction operation on V refers to the given ordering.

Assumption 2 is satisfied in most of the cases. Indeed, if, for a certain triplet of consecutive vertices, $d_{i-2,i}$ were perfectly equal to $d_{i-2,i-1} + d_{i-1,i}$, then the corresponding three atoms would be perfectly aligned. The Lebesgue measure of the subset

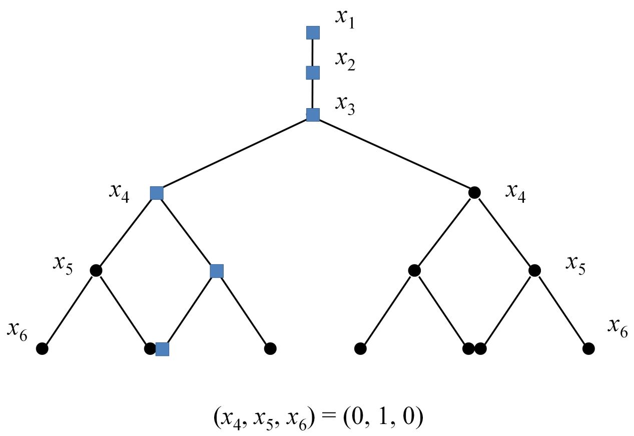

Fig. 2 Two representations for the solutions to the DMDGP: a path on the binary tree of atomic positions and a vector of $n-3$ binary variables.

of X not satisfying Assumption 2 is zero, so the probability of Assumption 2 not being satisfied is zero in a purely technical sense. Assumption 1 is instead harder to be satisfied. If data from NMR are considered, then only the distances smaller than 6Å are available, and therefore, if some of the distances $d_{i-3,i}$, $d_{i-2,i}$ and $d_{i-1,i}$ are large, then it is unknown and Assumption 1 may not be satisfied.

If both assumptions are satisfied and the atoms are positioned by following the same ordering given to the vertices of V, then only two possible positions can be chosen for the generic atom x_i. This property leads to the definition of a binary tree of possible atomic positions, where solutions to the DMDGP can be searched [12,18].

Referring to the binary tree of atomic positions, a solution can be seen as a complete path from the root of the tree to one of its leaf nodes. Moreover, a classic way to represent these solutions is through vectors of $n-3$ binary variables, where the i^{th} variable indicates which of the two possible choices for the atom x_i have been taken and $n = |V|$ (see Figure 2). Finally, note that a vector of $n-3$ binary variables can be seen as an integer number ranging from 0 to $2^{n-3}-1$. In this representation, the whole variability of the search domain is represented by a set of consecutive integer numbers. This representation is very useful for visualizing, for example, the LDE function (2), because it allows us to concentrate the variability of the search domain on only one coordinate axis (see Figure 3). The figure shows that the LDE function has several local minima in which a method for optimization could get stuck at. In the considered instance, there are only two molecular conformations corresponding to an LDE value equal to 0.

The BP algorithm [18] is an algorithm for the DMDGP which is based on the binary tree of solutions. The search proceeds by placing the atoms of the molecular conformation one per time: at each step, two possible positions for the current atom can be computed. As soon as an atomic position is found to be infeasible, then it is pruned and the search is backtracked. Pruning tests are employed for this purpose. Once the two possible positions for the current atom have been computed, the pruning test verifies whether the newly computed positions satisfy or not the available distances.

Algorithm 1 BP algorithm

```
0: BP($i$, $n$, $d$)
  for ($k$ = 1, 2) do
    compute the $k^{th}$ atomic position for the $i^{th}$ atom: $x_i^{(k)}$;
    check the feasibility of the atomic position $x_i^{(k)}$:
    if (the atomic position $x_i^{(k)}$ is feasible) then
      if ($i = n$) then
        one solution is found;
      else
        BP($i+1$,$n$,$d$);
      end if
    else
      the current branch is pruned;
    end if
  end for
```

If all the distances are satisfied (for a certain tolerance $\varepsilon > 0$), then the atomic position is feasible. If at least one of the distances is not satisfied, then the atomic position is not feasible. Because of the pruning phase, the size of the tree is reduced quickly and therefore an exhaustive search on the remaining branches is not too computationally expensive.

Algorithm 1 provides a sketch of the BP algorithm. The algorithm is invoked recursively, starting from the atomic position 4. The input parameters are i, the current atom whose position is searched; n, the total number of atoms; d, the set of known distances. A solution to the problem is found when BP(n,n,d) finds one feasible position for the last atom of the conformation.

We showed in previous works [12,14–18,25] that the BP algorithm is able to efficiently solve instances of the DMDGP. Theoretically speaking, atomic positions are only accepted in the BP if they are feasible, i.e. if the objective function is zero. Computationally, equality tests are replaced by ε-approximation tests, and hence some discrepancy from the zero cost is possible. It is important to note that, even though it is able to find solutions of a combinatorial optimization problem, the BP algorithm does not exploit any objective function. When the final set of solutions is found by BP,

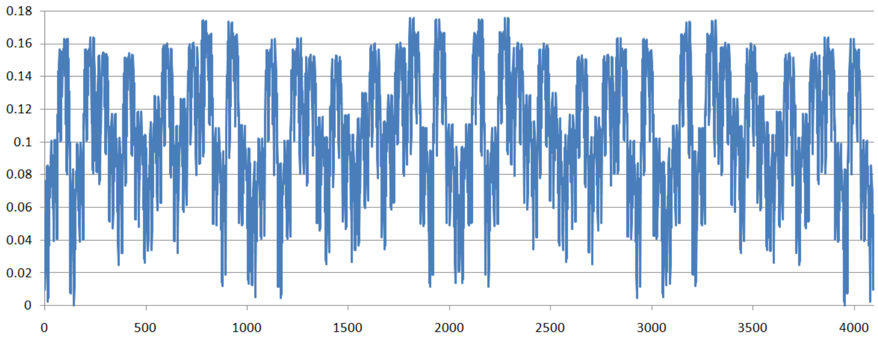

Fig. 3 A plot of the LDE function in correspondence with an instance of the DMDGP with 15 atoms.

the quality of the solutions can be evaluated through, for example, the LDE function (2).

3 Defining an artificial backbone of hydrogens

Let us suppose that the sequence of atoms N$-$C$_\alpha-$C (defining the protein backbones) and all the hydrogens H which are bound to such atoms are considered. Let $G = (V, E, d)$ be the associated weighted undirected graph. In order to reformulate the problem as a combinatorial problem, G must satisfy the two necessary assumptions. In particular, there cannot be triplets of aligned atoms, but the probability of having exactly three collinear atoms in any order is zero, so we shall assume that all possible orders satisfy Assumption 2. Moreover, Assumption 1 requires, given an ordering on V, that, for each atom x_i, there are three edges in E that precede i and that are incident in the vertex i. In other words, for each i, all the distances $d_{i-3,i}$, $d_{i-2,i}$ and $d_{i-1,i}$ must be known. Unfortunately, not all the vertices i of a graph G related to protein backbones can satisfy this assumption [27]. Therefore, the atoms related to these vertices need to be removed from G.

Note that some of the distances d in G can be computed by observations on the chemical structure of the protein backbone. When two atoms bind together, they form a known local conformation that is simply defined by the distance between the two atoms (which strongly depends on the kinds of bound atoms). Moreover, when three atoms bind as in a chain, where the first atom is bound to the second one, and the second atom is bound to the third one, we also obtain a local known conformation for the three atoms. As already remarked, the distances between the two pairs of bound atoms are known. Furthermore, the three atoms also form an angle that only depends by the kinds of the three interacting atoms. By exploiting the known distances and the value for the angle, the distance between the first and the third atom can be easily computed. To sum up, all the distances between bound atoms and between atoms which are bound to a common atom are known a priori. For example, the distances $d_{i,i+1}$ and $d_{i,i+2}$ related to the sequence of atoms N$-$C$_\alpha-$C are all known. However, in practice, all the distances that are known a priori are not enough for satisfying Assumption 1.

Therefore, only atoms whose relative distances can be obtained by NMR are exploited for having the combinatorial reformulation. As a consequence, only hydrogen atoms will be considered in the following (and hence only relative distances shorter than 6Å). Let us consider the subgraph G_H, such that $G \supset G_H = (V_H, E_H, d_H)$ and such that G_H contains only hydrogen atoms. We will refer to the set of atoms related to G_H as *artificial backbone* of hydrogens. The problem we need to solve is the one of identifying an artificial backbone of hydrogens such that the distances $d_{i-3,i}$, $d_{i-2,i}$ and $d_{i-1,i}$ are all smaller than 6Å and, as a consequence, detectable by NMR. In order to avoid confusion, the sequence of atoms N$-$C$_\alpha-$C will be referred to as *real backbone* of the protein.

There are 4 hydrogens in the common part of each amino acid. However, during the protein synthesis, consecutive amino acids bind to each other through a peptide bond. During this process, one of the hydrogens bound to the nitrogen N and the group OH bound to C separate from the other atoms and form a water molecule (H_2O) [29]. Therefore, the common part of each amino acid in a protein contains two hydrogens only (see Figure 1).

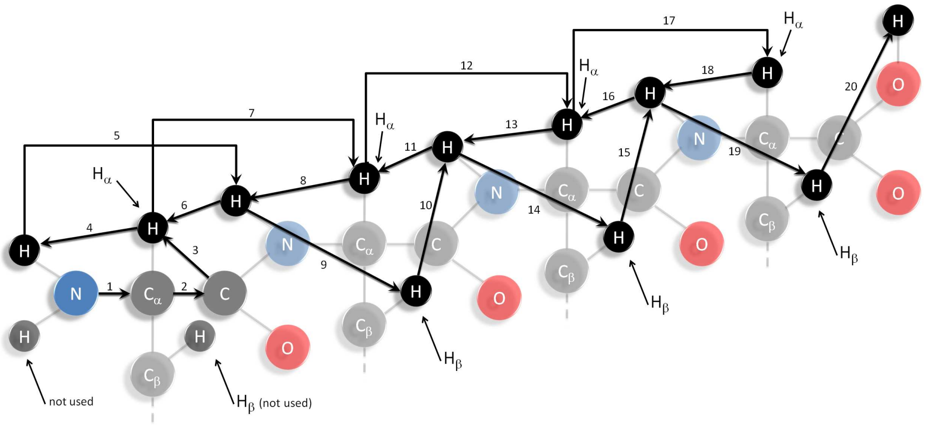

Fig. 4 The artificial backbone providing an ordering such that the assumptions for the DMDGP are satisfied (4 amino acids are considered). Note that some of the hydrogens are considered twice and that the considered order is specified through the labels associated to the arrows.

We will refer to the hydrogen bound to N with the symbol H, and we will refer to the hydrogen bound to C$_\alpha$ with the symbol H$_\alpha$. We also consider a third hydrogen which is not contained into the common part of each amino acid. It is borrowed from the *side chains* of the amino acids. Each amino acid has a different side chain: 19 of the 20 amino acids involved in the protein synthesis have a carbon atom C$_\beta$ in their side chains which is bound to the carbon atom C$_\alpha$. At least one hydrogen is bound to the carbon atom C$_\beta$, and we can consider one of them in the artificial backbone. The only exception is given by *glycine*, whose side chain consists in only one hydrogen atom. In the particular case of *glycine*, we consider as third hydrogen the only one that forms its side chain. We will refer to this third hydrogen with the symbol H$_\beta$.

The artificial backbone we consider is the one in Figure 4. A label is associated to each arrow for specifying the ordering given to the hydrogens. As the figure shows, the artificial backbone considers some of the hydrogens more than once, in order to bring the distance between the hydrogens $\{x_{i-3}, x_{i-2}, x_{i-1}, x_i\}$ under the 6Å threshold. Note that the nitrogen atom N, the carbon atom C$_\alpha$ and the carbon atom C of the first amino acid are also included in the ordering. These three atoms are also considered in the experiments together with the artificial backbone (see Section 5). We decided to add these three atoms at the beginning of our artificial backbone because they define a common coordinate system for the hydrogens and for the atoms of the real backbone.

The artificial backbones of hydrogens have some interesting properties. Since some of the hydrogens are considered twice, some of the relative distances between them are perfectly zero. If one of the distances between the atoms in the generic triplet $\{x_{i-2}, x_{i-1}, x_i\}$ is zero, then two atoms coincide and, as a consequence, the atoms of the triplet lie on the same straight line (this goes against Assumption 2). For this reason, the artificial backbone is built in a way that only distances d_{ij}, with $j > i + 2$, can be zero.

In the following, we will show that the artificial backbone of hydrogens can satisfy the necessary assumptions for the DMDGP (see Theorem 1). Then, we will also show that this artificial backbone can be extended by adding the atoms N, C_α and C of the first amino acid at the beginning (see Theorem 2), while the needed assumptions are still satisfied.

In the theorems, the superscripts i will indicate to which amino acid a given atom belongs. Therefore, N^i, C^i_α and C^i will refer to the three atoms of the real backbone of the i^{th} amino acid (the amino acids are sorted as in the sequence defining the protein). H^i will refer to hydrogen bound to N^i, H^i_α to the hydrogen bound to C^i_α, and H^i_β to the hydrogen taken from the side chain of the i^{th} amino acid. Finally, $H^{n'}$ will refer to the last hydrogen of the artificial backbone (see Figure 4), the only one that is bound to an oxygen. We can state the following theorem.

Theorem 1 *If all the distances*

$$d(H^1,H^1_\alpha), d(H^1,H^2), d(H^1,H^2_\alpha), d(H^1_\alpha,H^2), d(H^1_\alpha,H^2_\alpha), d(H^1_\alpha,H^2_\beta),$$

$$\forall i \in \{2,\dots,n-1\} \quad d(H^i,H^{i-1}_\alpha), d(H^i,H^i_\alpha), d(H^i,H^{i-1}_\beta), d(H^i,H^i_\beta), d(H^i,H^{i+1}),$$

$$\forall i \in \{2,\dots,n-1\} \quad d(H^i_\alpha,H^{i-1}_\beta), d(H^i_\alpha,H^i_\beta), d(H^i_\alpha,H^{i+1}_\beta), d(H^i_\alpha,H^{i+1}), d(H^i_\alpha,H^{i+1}_\alpha),$$

$$d(H^n,H^{n-1}_\alpha), d(H^n,H^n_\alpha), d(H^n,H^{n-1}_\beta), d(H^n,H^n_\beta), d(H^n,H^{n'}),$$

$$d(H^n_\alpha,H^n_\beta), d(H^n_\alpha,H^{n'}), d(H^n_\beta,H^{n'})$$

are obtained by NMR, then the artificial backbone defined as

$$H^1_\alpha, H^1, H^2, H^1_\alpha, \dots, H^i_\alpha, H^i, H^i_\beta, H^{i+1}, H^i_\alpha, \dots, H^n_\alpha, H^n, H^n_\beta, H^{n'}$$

satisfies Assumption 1 of the DMDGP.

Proof Let us consider the first 4 atoms of the artificial backbone:

$$(H^1_\alpha, H^1, H^2, H^1_\alpha).$$

This is a clique if all the relative distances between all the pairs of these atoms are known. All these distances are known by the hypothesis.

Let us now consider a generic quintuplet of consecutive hydrogens of the artificial backbone, given by

$$H^i_\alpha, H^i, H^i_\beta, H^{i+1}, H^i_\alpha,$$

for each $i \in \{2,\dots,n-1\}$. We will prove that, for each of these hydrogens h, the set of atoms formed by h and the three preceding ones forms a clique. To do this, we will show that, for each considered hydrogen h, the distances between the three preceding atoms and h are known.

Let us consider the generic hydrogen H^i. The three preceding atoms are H^i_α, H^{i-1}_α and H^i, ordered from the closest to the farthest atom. The distance between the two copies of H^i in the order (see discussion before statement of Theorem 1) is obviously zero. The distances $d(H^i,H^{i-1}_\alpha)$ and $d(H^i,H^i_\alpha)$ are both contained in the list of distances given in the theorem statement. Therefore, the generic hydrogen H^i of the artificial backbone satisfies Assumption 1.

The same observations can be made in order to prove that each atom of a generic quintuplet of consecutive hydrogens of the artificial backbone satisfies Assumption 1. Note that, when considering the generic hydrogen H^i_α, one must be aware that there are two atoms labeled in this way in the artificial backbone, but they are in two different positions in the sequence. Depending on which of the two is chosen, the three preceding atoms are obviously different.

Finally, let us consider the last four atoms of the artificial backbone:

$$(\mathrm{H}^n_\alpha, \mathrm{H}^n, \mathrm{H}^n_\beta, \mathrm{H}^{n'}).$$

Since all the distances between these atoms and the three preceding ones are known by the hypothesis, the theorem is proved. □

Since NMR experiments can find estimates of distances shorter than 6Å, we have the following immediate corollary:

Corollary 1 *If all the distances in the hypothesis of Theorem 1 are smaller than 6Å, then the corresponding artificial backbone of hydrogens satisfies Assumption 1.*

In the computational experiments in Section 5, for each considered artificial backbone, all the distances in the hypothesis of Theorem 1 are smaller than 6Å. Therefore, the combinatorial reformulation is possible for these instances, and the BP algorithm is able to efficiently find solutions related to the artificial backbones. However, in practice, few of these distances could be larger than 6Å. In order to verify this, we considered the distances available from the Cambridge Structural Database [7] for the bond lengths and bond angles among atoms, and we made some computations for verifying which distances could be out of the 6Å threshold. Our computations show that, among all the distances in the hypothesis of Theorem 1, there are only 3 of them that could exceed the threshold.

The first one is the distance between H^1 and H^2_α, whose maximum is 6.57Å. This distance does not pose any real trouble, because it is the distance between a hydrogen of the first amino acid and another hydrogen of the second amino acid. Hence, it appears only once. In the event that this distance should be larger than 6Å, one way to solve the problem could be to investigate all the possible orderings of the first few atoms of the artificial backbone in order to have Assumption 1 satisfied. This procedure is not expensive and it is easily implementable.

The other two distances that could not be detected by NMR are:

$$d(\mathrm{H}^i_\alpha, \mathrm{H}^{i-1}_\beta) \quad \text{and} \quad d(\mathrm{H}^i_\alpha, \mathrm{H}^{i+1}_\beta),$$

whose maximum values are 6.65Å and 6.61Å, respectively. In these two cases, another hydrogen bound to C_β can be considered in order to shorten the relative distances between the hydrogens of the artificial backbone. Obviously, this procedure cannot be applied when the corresponding amino acid is a *glycine*. In the worst case, in which no orderings for the first atoms of artificial backbone can be found and no more hydrogens bound to C_β can be added, we can substitute the three distances greater than 6Å by the three intervals:

$$d(\mathrm{H}^1, \mathrm{H}^2_\alpha) \in [6.00, 6.57], \quad d(\mathrm{H}^i_\alpha, \mathrm{H}^{i-1}_\beta) \in [6.00, 6.65], \quad d(\mathrm{H}^i_\alpha, \mathrm{H}^{i+1}_\beta) \in [6.00, 6.61].$$

An extension of the BP algorithm which is able to manage intervals instead of exact distances is currently under study. Preliminary results have been presented in [23].

Note that, since some of the distances between the hydrogens of the artificial backbones are zero, the LDE function (2) cannot be used for evaluating the quality of the solutions found by the BP algorithm in correspondence with the considered instances, because there would be divisions by zero. Therefore, in the experiments showed in the next section, we will consider a modified version of the LDE function, in which the terms that would contain divisions by zero are discarded.

As already mentioned, we consider in our computational experiences an extension of the artificial backbone defined in Theorem 1, where the atoms N, C_α and C of the first amino acid are added at the beginning (see Figure 4). We can state the following theorem.

Theorem 2 *If the artificial backbone defined in Theorem 1 satisfies Assumption 1 of the DMDGP, then the sequence of atoms*

$$\mathrm{N}, \mathrm{C}_\alpha, \mathrm{C}, \mathrm{H}^1_\alpha, \mathrm{H}^1, \mathrm{H}^2, \mathrm{H}^1_\alpha, \ldots, \mathrm{H}^i_\alpha, \mathrm{H}^i, \mathrm{H}^i_\beta, \mathrm{H}^{i+1}, \mathrm{H}^i_\alpha, \ldots, \mathrm{H}^n_\alpha, \mathrm{H}^n, \mathrm{H}^n_\beta, \mathrm{H}^{n'},$$

obtained by adding the atoms N, C_α *and* C *at the beginning of the artificial backbone, also satisfies Assumption 1 of the DMDGP.*

Proof In order to have Assumption 1 satisfied for the added atoms, we need to verify that these three subsets of atoms

$$(\mathrm{N}, \mathrm{C}_\alpha, \mathrm{C}, \mathrm{H}^1_\alpha), \quad (\mathrm{C}_\alpha, \mathrm{C}, \mathrm{H}^1_\alpha, \mathrm{H}^1), \quad (\mathrm{C}, \mathrm{H}^1_\alpha, \mathrm{H}^1, \mathrm{H}^2),$$

form three cliques. In the first subset $(\mathrm{N}, \mathrm{C}_\alpha, \mathrm{C}, \mathrm{H}^1_\alpha)$, the distances $d(\mathrm{N}, \mathrm{C}_\alpha)$, $d(\mathrm{N}, \mathrm{C})$, $d(\mathrm{C}_\alpha, \mathrm{C})$, $d(\mathrm{C}_\alpha, \mathrm{H}^1_\alpha)$ and $d(\mathrm{C}, \mathrm{H}^1_\alpha)$ can be obtained by using the bond lengths and bond angles between these atoms, which are known a priori. Moreover, by the symmetric property of the DMDGP [12], we can also fix the fourth atom (in this case H^1_α), in order to obtain the distance $d(\mathrm{N}, \mathrm{H}^1_\alpha)$.

Finally, all the needed distances in the other two subsets $(\mathrm{C}_\alpha, \mathrm{C}, \mathrm{H}^1_\alpha, \mathrm{H}^1)$ and $(\mathrm{C}, \mathrm{H}^1_\alpha, \mathrm{H}^1, \mathrm{H}^2)$ can be obtained by using the known bond lengths and bond angles. □

4 Reconstructing the real backbone of a protein conformation

Let us suppose that the coordinates of the hydrogens of an artificial backbone have been obtained by the BP algorithm. This information can be used in order to identify all the other atoms of the protein backbone, i.e. the sequence of atoms $\mathrm{N} - \mathrm{C}_\alpha - \mathrm{C}$. One way for doing this is to exploit the available coordinates of the hydrogens and the distances (known a priori) between these hydrogens and other atoms of the protein backbone.

The procedure we use for building the real backbone of a protein from an artificial backbone is the following. Let us suppose that we need to find the coordinates of the backbone atom a and that the distances between a and other four atoms b_1, b_2, b_3 and b_4 (with known coordinates) are known. In this hypothesis, the coordinates of the atom a can be identified if the 4 atoms are non-coplanar [6,31,32].

Let d_{a,b_i} be the Euclidean distance between the atom a and the atom b_i, for all $i \in \{1, 2, 3, 4\}$. In order to find the coordinates of a, the following system needs to be solved:

$$\begin{cases} ||a - b_1|| = d_{a,b_1} \\ ||a - b_2|| = d_{a,b_2} \\ ||a - b_3|| = d_{a,b_3} \\ ||a - b_4|| = d_{a,b_4}. \end{cases} \tag{3}$$

This is a quadratic system of 4 equations in 3 variables.

Theorem 3 *If the quadratic system (3) has a solution, then the linear system*

$$Ax = b, \tag{4}$$

where

$$A = -2 \begin{bmatrix} 1 & (b_1)^T \\ 1 & (b_2)^T \\ 1 & (b_3)^T \\ 1 & (b_4)^T \end{bmatrix}, \quad x = \begin{pmatrix} t \\ a \end{pmatrix}, \quad b = \begin{bmatrix} \left(d^2_{a,b_1}\right) - \left(||b_1||^2\right) \\ \left(d^2_{a,b_2}\right) - \left(||b_2||^2\right) \\ \left(d^2_{a,b_3}\right) - \left(||b_3||^2\right) \\ \left(d^2_{a,b_4}\right) - \left(||b_4||^2\right) \end{bmatrix}$$

and

$$t = -\frac{||a||^2}{2},$$

has the same solution.

Proof Let us square both sides of the quadratic equations of system (3):

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

The independent variables in the quadratic terms in all the equations correspond to the coordinates of a. Therefore, in order to remove the quadratic terms, we can introduce another variable $t = -||a||^2/2$. The introduction of t allows to obtain a linear system of 4 equations in 4 variables, that can be written as system (4). So, if the quadratic system (3) has a solution, then the linear system (4) has the same solution. □

Note that the result of Theorem 3 is related to a result reported in [6]. However, in the quoted paper, the system (3) is transformed into a 3 × 3 linear system, where the three independent variables correspond to the three coordinates of the atom a. The advantage in using the 4 × 4 linear system (4) is that the introduced variable t allows to check the quality of the computed solutions. Indeed, when solving the system (4), one solution could be found even though the original system (3) has no solutions. This could happen when the distances d_{a,b_i} are not compatible to one another. If the value of t for a found solution corresponds to the squared norm of the vector a related to the remaining variables ($t = -||a||^2/2$), for a given tolerance, then the vector a can be considered as a good estimate of the coordinates of the considered atom.

Theorem 4 *The real backbone of a protein conformation can be calculated in linear time by using the artificial backbone defined in Theorem 1, with the atoms* N, C$_\alpha$, C *added at the beginning (see Theorem 2).*

Proof The positions of all the backbone atoms N, C$_\alpha$ and C can be found by solving a sequence of 4 × 4 linear systems defined by (4). Three different systems can be defined, depending on the atom whose coordinates need to be computed. For computing the position of the nitrogen N of the protein backbone, for example, the following four atoms with known positions can be considered: C$_\alpha$ and C of the previous amino acid, the hydrogen H bound to N and the hydrogen H$_\alpha$ bound to the following C$_\alpha$ (see Figure 5). The considered C$_\alpha$ and C are supposed to be known (if they belong to the first amino acid, they are known because they are the first three atoms (including N) that are fixed by BP considering the extension of the artificial backbone defined in Theorem 2). The distances between C and N and between N and H are known because these two pairs of atoms are chemically bound. The distance between C_α and N is also known, because the bond lengths $C_\alpha - C$ and $C - N$ are known, and the angle among the three atoms $C_\alpha - C - N$ is also known. For the same reason, the distance between N and H_α is available. The solution of the corresponding 4×4 linear system allows to identify the coordinates of N. Analogue observations can be made for the other two systems. See Figure 5 to find out which atoms and distances can be considered. □

5 Computational experiments

We will show in this section how instances of the DMDGP related to artificial backbones can be efficiently solved by the BP algorithm, and how the solutions found by BP can be exploited for reconstructing the real backbone of a protein conformation. All the codes were written in C programming language and all the experiments were carried out on an Intel Core 2 CPU 6400 @ 2.13 GHz with 4GB RAM, running Linux. The codes have been compiled by the GNU C compiler v.4.1.2 with the `-O3` flag.

The instances we consider are artificially generated. The method we use for generating such instances is similar to the one proposed in [13]. However, in this case, not only the backbone atoms N, C_α and C are considered, but also hydrogens H, H_α and H_β. The atoms of the real backbone are used only for placing the hydrogen atoms in a way that they simulate protein conformations, and they are discarded when creating the final instance. Some hydrogen atoms are considered twice and they are all sorted in accordance with the special ordering of the artificial backbone discussed in Section 3. Only distances smaller than 6Å are considered. We randomly generated a set of instances having a different number of amino acids. When the real backbone is reconstructed, each amino acid is represented by the 3 hydrogens H, H_α and H_β, the carbons C_α and C, and the nitrogen N.

All the instances we generated belong to the class of instances of the DMDGP. We applied the BP algorithm for solving such instances, and the computational experiments are shown in Table 1. In the table, n is the number of atoms included in the instance. It is always a number which is divisible by 5, because each amino acid of the considered artificial backbone contains exactly 5 hydrogens (two of them are considered twice). As a consequence, the number of amino acids in each instance is $m = \frac{n}{5}$, and the number of atoms (all distinct to each other) after the reconstruction of the real backbone will be $6m$. The cardinality of the set of edges E indicates the number

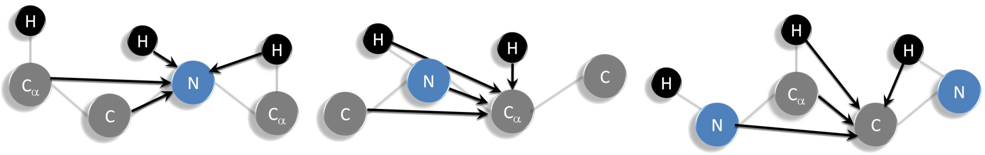

Fig. 5 The atoms and the distances used in the three linear systems used to determine the protein backbone.

of known distances, and #Sol is the number of found solutions. The LDE function (modified in order to avoid divisions by zero) is used for evaluating the quality of the solutions and the best one is showed in the table. Finally, the CPU time (in seconds) is given for each experiment. The experiments show that the BP algorithm is very efficient in finding solutions of the DMDGP in terms of quality of the solutions and CPU time, as already shown in previous works. In these experiments, each solution consists of the set of coordinates of the hydrogen atoms H, H_α and H_β of the artificial backbones. The total number #Sol of solutions is 2 or 4. When there are 4 solutions, the experiments are slightly more expensive because more branches of the binary tree are explored during the search.

By following the method described in Section 4, we also built the real backbone, i.e. the sequence of atoms $\mathrm{N} - \mathrm{C}_\alpha - \mathrm{C}$, related to the solutions found by BP. To this aim, a sequence of linear systems (4) needs to be solved, and we solved each of them by using the procedure `dgesv` of the LAPACK library [2]. Even though these linear systems are small in size and they are solved by one of the best solvers for linear systems, the corresponding solutions are very sensible to errors. In theory, the solution of the system (4) can provide the correct coordinates of the atom only if these coordinates are also solution for the original system (3). In practice, approximations of such coordinates can be found by solving system (4) even when floating-point errors affect the used distances. Unfortunately, approximated values of the coordinates can bring to the propagation of errors along the atoms of the real backbone, that can spoil the final real backbone which is obtained.

The strategy we implemented for overcoming this problem is the following. As explained in Section 4 (see Theorem 4), depending on the kind of real backbone atom we want to place, 4 different distances must be used in the linear system. Such distances are known, and we obtained them from the Cambridge Structural Database [7]. These distances can be considered as correct, but the floating-point arithmetic of a computer machine can make them incompatible with known atomic coordinates, even though this incompatibility can be caused by very small errors. Therefore, we tune the known distances to the current system to be solved by slightly modifying their values. In the experiments, each distance is allowed to change its values into a small neighbor of its known value. Many systems are then solved by using different combinations for the values for the distances, and the quality of the found solutions is checked through the variable t. A solution is accepted only if the difference between t and $-\frac{1}{2}||a||^2$ is smaller than a given small tolerance. In this way, we do not allow errors to propagate along the

Instance name	n	$\|E\|$	#Sol	LDE	CPU time
rand1	50	444	4	1.75e-09	0.00
rand2	100	1180	2	3.42e-09	0.00
rand3	200	2872	2	1.00e-08	0.01
rand4	400	5867	2	9.70e-09	0.01
rand5	800	13460	2	1.40e-08	0.03
rand6	1500	22040	4	9.13e-09	0.14
rand7	3000	54537	4	6.43e-08	0.43
rand8	5000	87992	2	2.35e-08	0.80

Table 1 Results obtained by the BP algorithm on a set of randomly generated artificial backbones.

Instance name	$6m$	#Sol	*err*
rand1	60	4	0.000021
rand2	120	2	0.000046
rand3	240	2	0.000408
rand4	480	2	0.008775
rand5	960	2	0.001793
rand6	1800	4	0.005795
rand7	3600	4	0.012779
rand8	6000	2	0.011418

Table 2 The *err* value for the best real backbone reconstructed by solving the sequence of 4×4 linear systems for each instance in Table 1.

atoms of the real backbone, and we are able to find very accurate positions for these atoms.

We computed the real backbone of all the solutions found by BP in correspondence with the instances in Table 1. Table 2 shows the obtained results. $6m$ is the number of atoms of the final conformation, after the real backbone has been built. #Sol is the number of solutions found by BP: we reconstructed the real backbone for all of them, and the one with the best average error is considered in the table. The quality of the found real backbones is evaluated through the function:

$$err = \frac{1}{3m} \sum_{i=1}^{3m} | \, t + \frac{||a||^2}{2} \, |,$$

which computes an average of all the single errors occurring while solving the linear systems. Note that this function depends on the solutions of the linear systems only (we have information on the actual coordinates of the atoms because the instances are artificially generated, but they are not used here). Since three backbone atoms are computed for each amino acid, the average is computed on $3m$ local errors. We can see from Table 2 that the value of *err* is approximately 10^{-3}, and it is close to 10^{-2} in correspondence with larger instances. This indicates that the propagation of the errors through the linear systems has been kept low (some of the known distances have a precision of 10^{-3}), and that good quality real backbones have been reconstructed.

6 Conclusions

In this paper we solve the MDGP on protein backbones of up to 6000 atoms. The main innovation we put forward is to take into account a limitation of NMR measurements, i.e. the fact that interatomic distances are usually only obtained for hydrogen atoms (this limitation is usually overlooked in the literature). We address it by identifying an “artificial backbone of hydrogens” with two properties: (i) it consists of hydrogen atoms only; (ii) it is an instance of the DMDGP, and so it can be solved by a very efficient discrete search method called Branch & Prune. Once the hydrogen backbone is positioned, its embedding is extended to the other atoms in the natural backbone by solving a few (small) linear systems.

Future works will be devoted to extensions of the presented strategies for managing the noise and the errors that can affect experimentally obtained data. The distances between the hydrogens that are given to the BP algorithm may not be so accurate, and hence intervals in which the distances are contained might be available, instead of exact distances. Moreover, a few of these distances could be wrong. We plan to combine all these ideas into a general strategy, and to extend such a strategy to entire protein conformations.

Acknowledgments

The authors wish to thank the anonymous referees, whose comments helped us improving this paper. The authors would also like to thank the Brazilian research agencies FAPESP and CNPq, the French research agency CNRS and École Polytechnique, for financial support.

References

1. L.T.H. An, P.D. Tao, *Large-Scale Molecular Optimization from Distance Matrices by a D.C. Optimization Approach*, SIAM Journal on Optimization **14**, 77–114, 2003.
2. E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3^{rd} edition, SIAM, 1999.
3. 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**, 1251–1277, 2008.
4. R.S. Carvalho, C. Lavor, and F. Protti, *Extending the Geometric Buildup Algorithm for the Molecular Distance Geometry Problem*, Information Processing Letters **108**, 234–237, 2008.
5. G.M. Crippen and T.F. Havel, *Distance Geometry and Molecular Conformation*, John Wiley & Sons, New York, 1988.
6. Q. Dong, 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.
7. R.A. Engh, R. Huber, *Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement*, Acta Crystallographica **A47**, 392–400, 1991.
8. A. Grosso, M. Locatelli, F. Schoen, *Solving Molecular Distance Geometry Problems by Global Optimization Algorithms*, Computational Optimization and Applications **43**, 23–27, 2009.
9. T.F. Havel, *Distance Geometry*, D.M. Grant and R.K. Harris (Eds.), Encyclopedia of Nuclear Magnetic Resonance, Wiley, New York, 1701-1710, 1995.
10. S. Kirkpatrick, C.D. Jr. Gelatt and M.P. Vecchi, *Optimization by Simulated Annealing*, Science **220**(4598), 671–680, 1983.
11. C. Lavor, L. Liberti, and N. Maculan, *Molecular Distance Geometry Problem*, In: Encyclopedia of Optimization, C. Floudas and P. Pardalos (Eds.), 2^{nd} edition, Springer, New York, 2305–2311, 2009.
12. C. Lavor, L. Liberti, and N. Maculan, *Discretizable Molecular Distance Geometry Problem*, Tech. Rep. q-bio.BM/0608012, arXiv, 2006.
13. C. Lavor, *On generating Instances for the Molecular Distance Geometry Problem*, In: Global Optimization From Theory to Implementation, Leo Liberti and Nelson Maculan (Eds.), Series: Nonconvex Optimization and Its Applications **84**, Springer, 405–414, 2006.
14. C. Lavor, L. Liberti, A. Mucherino, and N. Maculan, *On a Discretizable Subclass of Instances of the Molecular Distance Geometry Problem*, ACM Conference Proceedings, 24^{th} Annual ACM Symposium on Applied Computing (SAC09), Hawaii USA, 804–805, 2009.
15. C. Lavor, A. Mucherino, L. Liberti, and N. Maculan, *Computing Artificial Backbones of Hydrogen Atoms in order to Discover Protein Backbones*, IEEE Conference Proceedings, International Conference IMCSIT09, Workshop on Combinatorial Optimization (WCO09), Poland, 751–756, 2009.

16. C. Lavor, A. Mucherino, L. Liberti, and N. Maculan, *An Artificial Backbone of Hydrogens for Finding the Conformation of Protein Molecules*, IEEE Conference Proceedings, Computational Structural Bioinformatics Workshop (CSBW09), Washington D.C., USA, 152–155, 2009.
17. C. Lavor, A. Mucherino, L. Liberti, and N. Maculan, *Discrete Approaches for Solving Molecular Distance Geometry Problems using NMR Data*, to appear in International Journal of Computational Biosciences, 2010.
18. 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), 1–17, 2008.
19. L. Liberti, C. Lavor, N. Maculan, F. Marinelli, *Double Variable Neighbourhood Search with Smoothing for the Molecular Distance Geometry Problem*, Journal of Global Optimization **43**, 207–218, 2009.
20. L. Liberti, C. Lavor, A. Mucherino, N. Maculan, *Molecular Distance Geometry Methods: from Continuous to Discrete*, to appear in International Transactions in Operational Research, 2010.
21. J.J. Moré, Z. Wu, *Global Continuation for Distance Geometry Problems*, SIAM Journal on Optimization **7**, 814–836, 1997.
22. J.J. Moré, Z. Wu, *Distance Geometry Optimization for Protein Structures*, Journal of Global Optimization **15**, 219–234, 1999.
23. A. Mucherino, C. Lavor, *The Branch and Prune Algorithm for the Molecular Distance Geometry Problem with Inexact Distances*, World Academy of Science, Engineering and Technology (WASET), Proceedings of the "International Conference on Bioinformatics and Biomedicine" (ICBB09), Venice, Italy, October 2009.
24. A. Mucherino, L. Liberti, C. Lavor, `MD-jeep`*: an Implementation of a Branch & Prune Algorithm for Distance Geometry Problems*, Lectures Notes in Computer Science, Proceedings of the Third International Congress on Mathematical Software (ICMS10), Kobe, Japan, September 2010.
25. A. Mucherino, C. Lavor, and N. Maculan, *The Molecular Distance Geometry Problem Applied to Protein Conformations*, Proceedings of the 8^{th} Cologne-Twente Workshop on Graphs and Combinatorial Optimization (CTW09), S. Cafieri, A. Mucherino, G. Nannicini, F. Tarissan, L. Liberti (Eds.), 337–340, Paris, 2009.
26. A. Mucherino, L. Liberti, C. Lavor, and N. Maculan, *Comparisons between an Exact and a MetaHeuristic Algorithm for the Molecular Distance Geometry Problem*, ACM Conference Proceedings, Genetic and Evolutionary Computation Conference (GECCO09), Montréal, Canada, 333–340, 2009.
27. A. Mucherino, C. Lavor, L. Liberti, and N. Maculan, *On the Definition of Artificial Backbones for the Discretizable Molecular Distance Geometry Problem*, Mathematica Balkanica **23**(3-4), 289-302, 2009.
28. J.B. Saxe, *Embeddability of Weighted Graphs in k-space is Strongly NP-hard*, Proceedings of 17^{th} Allerton Conference in Communications, Control, and Computing, Monticello, IL, 480–489, 1979.
29. T. Schlick, *Molecular Modelling and Simulation: an Interdisciplinary Guide*, Springer, New York, 2002.
30. C.D. Schwieters, J.J. Kuszewski, G.M. Clore, *Using Xplor-NIH for NMR Molecular Structure Determination*, Progress in Nuclear Magnetic Resonance Spectroscopy **48**, 47–62, 2006.
31. 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.
32. D. Wu, Z. Wu, Y. Yuan, *Rigid Versus Unique Determination of Protein Structures with Geometric Buildup*, Optimization Letters **2**, 319–331, 2008.
33. H. Xu, S. Izrailev, D.K. Agrafiotis, *Conformational Sampling by Self-Organization*, Journal of Chemical Information and Computer Sciences **43**, 1186–1191, 2003.