On the solution of molecular distance geometry problems with interval data

C. Lavor* L. Liberti† and A. Mucherino‡

*IMECC-UNICAMP, Campinas-SP, Brazil.
clavor@ime.unicamp.br
†LIX, École Polytechnique, Palaiseau, France.
liberti@lix.polytechnique.fr
‡CERFACS, Toulouse, France.
mucherino@cerfacs.fr

***Abstract*—The Molecular Distance Geometry Problem consists in finding the three-dimensional conformation of a protein using some of the distances between its atoms provided by experiments of Nuclear Magnetic Resonance. This is a continuous search problem that can be discretized under some assumptions on the known distances. We discuss the case where some of the distances are subject to uncertainty within a given nonnegative interval. We show that a discretization is still possible and propose an algorithm to solve the problem. Computational experiments on a set of artificially generated instances are presented.**

I. Introduction

In this paper we consider the problem of finding the three-dimensional conformation (the coordinates of all the atoms) of a protein from a subset of inter-atomic distances found using Nuclear Magnetic Resonance (NMR) experiments. This problem is usually referred to as Molecular Distance Geometry Problem (MDGP) [5], [10].

Proteins are important molecules which perform several functions in living beings. If their three-dimensional conformations are discovered, they are able to reveal the specific function that each protein is supposed to perform. A web database named Protein Data Bank (PDB) [1] is collecting all the three-dimensional conformations of proteins that scientists in the world have been able to obtain. To date, a large percentage of conformations on the PDB have been obtained through NMR experiments, where the corresponding MDGP has been solved by general-purpose continuous approaches for global optimization. The meta-heuristic Simulated Annealing [3], [14] is employed in most of the cases.

Since we focus our attention on proteins and on NMR experiments for obtaining estimates of inter-atomic distances, we are able to make the following assumptions, which will allow us to discretize the problem: 1) Inter-atomic distances corresponding to the set of all (unordered) pairs of atoms separated by at most two covalent bonds will be represented by positive rational numbers which will be held fixed, since they can be considered fixed in the majority of the protein conformation calculations [15]; 2) The pairs of atoms separated by exactly three covalent bonds, for which it is possible to compute tight lower and upper bounds to the corresponding distances, will be represented by intervals of rational numbers and the possible values will be represented by a discretized set of values within this interval.

In general, the solution of the MDGP requires a continuous search [10]. In this paper, we discretize the problem and we propose an extension of the Branch-and-Prune (BP) algorithm, given in [9], in order to consider uncertainties on the given distances. The distances from NMR are only used for pruning purposes. As a consequence, the new discrete domain of the problem is completely independent from the instance to be solved, and it cannot be spoiled by wrong data which might be obtained experimentally. As already remarked in our previous publications, the advantages in considering a discrete search, with respect to a continuous one, are: increased efficiency, increased solution accuracy and completeness (in the sense that all solutions can be found).

The discretization of the search space is based on the observation that, in general, three spheres in $\mathbb{R}^3$ intersect in at most two points. A technique for reliably computing such intersection points is given in [2].

Our first attempt to consider NMR data, which usually provide just distances between hydrogen atoms closer than a given threshold, has been presented in [6], [7], [12]. We defined an ordering for the hydrogens related to protein backbones which allows us to have information enough to perform the discretization. The black arrows and their labels in Figure 1 show the particular considered ordering. In [8], we proved that, because of steric constraints due to the particular structure of protein backbones, all the distances which are required for the discretization are smaller than 6Å. As a consequence, since NMR experiments are usually able to provide distances that are shorter than 6Å, all the necessary distances should be available.

Even though we showed that this approach works on a set of artificially generated instances, we remarked its limitations when we tried to apply it to real NMR data. The main issue is that the distances obtained by NMR are not precise, and they can be, in general, represented by intervals. As discussed in [13], discretizing with interval data leads to the computation of three spherical shells (instead of three spheres), which is quite

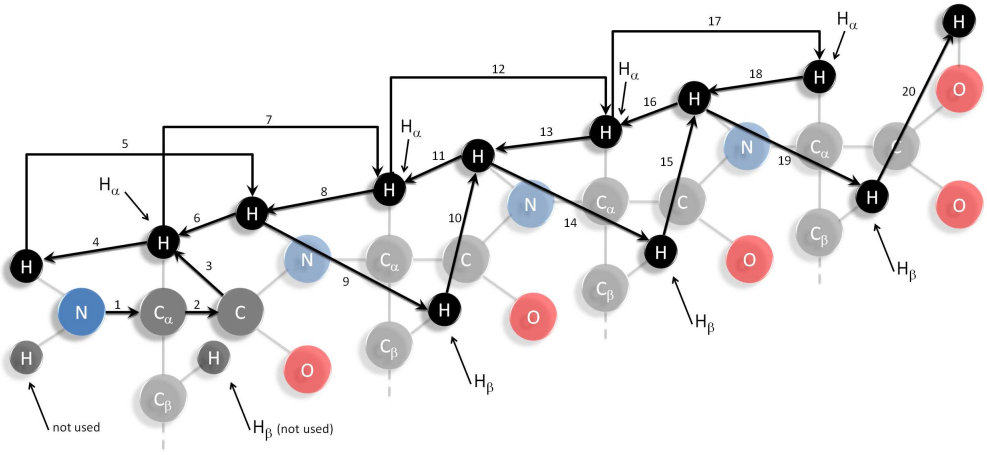

Fig. 1. Note that some of the hydrogens are considered twice and that the considered order is specified through the labels associated to the arrows.

complex to compute. We will introduce a different strategy in this paper, which is, however, incompatible with the artificial backbone in Figure 1.

Our first attempt to consider interval data has been presented in [11]. We assumed that the distances provided by NMR experiments are defined by a lower and an upper bound, and we modified the pruning phase of the BP algorithm from a "by value" form to a "by interval" form. However, needed distances for the discretization were still supposed to be exact. We observed that even a very low uncertainty on these distances is able to spoil the discretization process and no solutions can be found.

In the present work, we remove the latter phenomenon. Even though interval data are used, we will be able to mantain the discretization process. Also, we compute the positions of non-hydrogen atoms by a different method, thereby avoiding the numerical instabilities due to solving linear systems, as it was done in [8]. The new algorithm relies on a carefully hand-crafted atom sequence which exploits repetitions in order to make sure that for each atom being placed there are distances to three previously placed atoms, and that these distances guarantee that discretization can occur independently of the presence of interval represented distances, and a particular instance of the MDGP.

The rest of this paper is organized as follows. In Sect. II we introduce notation, some main concepts, and give some preliminary definitions. In Sect. III we construct the protein backbone graph and a vertex sequence that allows the discretization of the search space. In Sect. IV we propose the algorithm for the protein backbone graph using the order in the vertex sequence and present some computational results. Sect. V concludes the paper.

II. The Discretizable Molecular Distance Geometry Problem

For a graph $G = (V, E)$ and a subset $V_0 \subseteq V$ we let $G[V_0]$ be the subgraph of G induced by V_0; for $v \in V$ we let $\delta_E(v) = \{u \in V \mid \{u, v\} \in E\}$ be the set of vertices adjacent to v (if there is no ambiguity we omit the E index). For an order $<$ on V and $v \in V$ we let $\gamma_<(v) = \{u \in V \mid u < v\}$ be the set of predecessors of v in the order $<$ and $\rho_<(v) = |\gamma_<(v)| + 1$ be the rank of v in the order $<$ (if there is no ambiguity we omit the $<$ index).

In [4], [9] we introduced a subclass of MDGP whose instances can be solved using a discrete search algorithm.

> Discretizable Molecular Distance Geometry Problem (DMDGP). Given a nonnegatively weighted graph $G = (V, E, d)$ where $d : E \to \mathbb{R}_+$, a subset $V_0 \subseteq V$ and an order $<$ on V such that:
>
> - $V_0 = \{1, 2, 3\}$ and $G[V_0]$ is a clique (Start)
> - for all $v \in V \setminus V_0$ we have
>   - $v-3, v-2, v-1 \in \delta(v) \cap \gamma(v)$ (Discretization)
>   - $d(v-3, v-2) + d(v-2, v-1) > d(v-3, v-1)$ (Strict Triangular Inequalities),
>
> is there an embedding $x : V \to \mathbb{R}^3$ such that
>
> $$\forall \{u, v\} \in E \quad \|x(u) - x(v)\| = d(u, v) \qquad (1)$$
>
> holds ?

The vertices of G correspond to the atoms forming the molecule and edges indicate if the distance between the respective atoms is known or not.

The DMDGP is **NP**-hard [4] and its instances can be solved using the BP algorithm [9]: the first 3 vertices in V_0 can be embedded by Start; inductively, any vertex v of rank greater than 3 can be placed at the intersection of three spheres centered at $v-3, v-2, v-1$ with respective radii $d(v-3, v)$, $d(v-2, v)$, $d(v-1, v)$ by Discretization; this intersection consists of at most 2 points x'_v, x''_v by Strict Triangular Inequalities. This gives rise to a binary tree search whose leaves represent valid embeddings of G. Branches can be pruned using distances from v to vertices in $\delta(v) \cap \gamma(v)$ (other than the ones used for the discretization) that are incompatible with either x'_v or x''_v or both. This yields an extremely fast

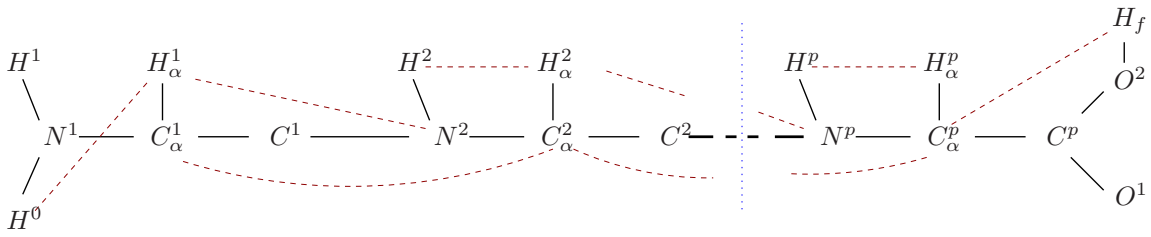

Fig. 2. A graph representing the general structure of a protein backbone. Dashed lines show some distances which need to be represented by intervals.

algorithm [9] which is also able to find all embeddings for a given graph (modulo rotations and translations).

In order to facilitate our task, we allow for repeated atoms in the ordering that we will define in the next section. This trick allows us to consider distances between copies of the same atom, that are naturally equal to 0, thus increasing the number of exact distances that can be considered. Obviously, since the same atom can be duplicated several times, the final sequence of atoms could have a length which is much larger than the original sequence of atoms. However, this increase in length is not reflected on the tree obtained by the discretization, because copies of an atom which has been already placed somewhere can only take one position. In other words, there is no branching on the tree in correspondence with duplicated atoms.

III. An artificial order for protein backbones

Figure 2 shows the general structure of a protein backbone, where superscripts indicate the amino acid to which each atom belongs. H^0 is the second hydrogen that is bond to the first nitrogen N^1: this is the only case in which two hydrogens are bound to the same atom. H_f belongs to the last amino acid, and it is bound to the second oxygen O^2.

The atoms of the protein backbone can be ordered into a natural way. For example, if the following ordering is considered (see Figure 2):

$$\{H^0,\ H^1,\ N^1,\ C^1_\alpha,\ H^1_\alpha,\ C^1,\ \ldots,\ H^i,\ N^i,\ C^i_\alpha,\ H^i_\alpha,\ C^i,\ \ldots,\ H^p,\ N^p,\ C^p_\alpha,\ H^p_\alpha,\ C^p,\ O^1,\ O^2,\ H_f\},$$

then it is easy to verify that the assumptions for the discretization are not satisfied. However, we discovered a particular ordering for these atoms which allows us to discretize even if interval data are considered.

First we define the finite sequence (see Fig. 3):

$$r^1 = (N^1, H^1, H^0, C^1_\alpha, N^1, H^1_\alpha, C^1_\alpha, C^1, N^2, C^2_\alpha),$$

related to the first amino acid of the protein backbone. Then, for a given $i \in \{2, \ldots, p-1\}$, we define the finite sequence (see Fig. 4):

$$r^i = (H^i, N^i, C^i_\alpha, H^i_\alpha, C^i, C^i_\alpha, N^{i+1}, C^i, C^{i+1}_\alpha),$$

related to the generic amino acid of the protein backbone. Finally, the finite sequence (see Fig. 5):

$$r^p = (H^p, N^p, C^p_\alpha, H^p_\alpha, C^p, C^p_\alpha, O^2, C^p, H_f),$$

shows the ordering for the last amino acid. We remark that since O^1 has known precise distances to C^p_α, C^p, O^2 its placement is not problematic.

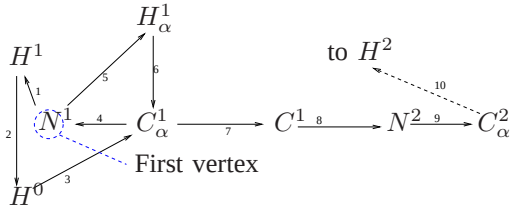

Fig. 3. The sequence r^1.

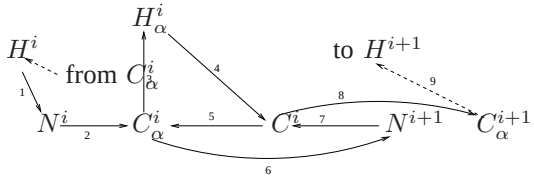

Fig. 4. A sequence r^i (for $i \in \{2, \ldots, p-1\}$).

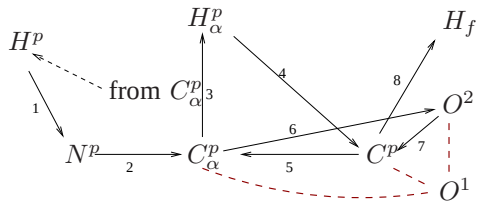

Fig. 5. The sequence r^p.

Thus, the sequence which defines a complete ordering for all the atoms of the protein backbone is:

$$r_{PB} = (r^1, r^2, \ldots, r^{p-1}, r^p).$$

We point out that the defined ordering allows to discretize MDGPs where NMR data are supposed to be represented by a set of intervals on the distances. Indeed, among the distances needed for the discretization, the distances $(i, i+1)$ and $(i, i+2)$ are always exact, because they are computed a priori by exploiting information on bond lenghts and angles. Only distances $(i, i+3)$ can be represented by intervals (they are marked by dashed lines in Figure 2). When this is the case,

the discretization process could be performed by computing the intersection of two spheres (related to exact distances) and a spherical shell (related to the interval). This procedure would be able to define a curve in the three-dimensional space in which the possible positions for the current atom can be searched. However, the equation of the curve would provide information on the atomic positions with a precision which is actually not needed for the purposes of the computation. Therefore, we discretize the interval related to the distance $(i, i+3)$ and apply the standard discretization process for a subset of sample distances extracted from the available interval.

IV. Computational results

We consider a very simple instance with 3 amino acids and a subset of instances with a larger number of amino acids. For the small instance containing only 3 amino acids, we analyze in details the defined ordering r_{PB}, given by

$$\begin{aligned} r_{PB} \quad = \quad & (N^1, H^1, H^0, C^1_\alpha, N^1, H^1_\alpha, C^1_\alpha, C^1, N^2, C^2_\alpha, \\ & H^2, N^2, C^2_\alpha, H^2_\alpha, C^2, C^2_\alpha, N^3, C^2, C^3_\alpha, \\ & H^3, N^3, C^3_\alpha, H^3_\alpha, C^3, C^3_\alpha, O^2, C^3, H_f). \end{aligned}$$

In order to keep a very high control on this first experiment, we consider only 3 different distances. In practice, every time we need a distance between two bound atoms, we always consider the same value d_1, independently from the kinds of atoms. Moreover, every time a distance between two atoms bound the a common atom is needed, the value d_2 is always considered. Finally, when the distance between two atoms separated by three chemical bonds is required, we consider the interval $[l_3, u_3]$, where l_3 is the minimum possible value and u_3 is the maximum value for the distance. The same values are repeated along the whole sequence.

The distances d_1 and d_2, as well as the interval $[l_3, u_3]$, provide the information which is needed for computing the discrete search domain. The associated tree contains all the possible solutions related to protein backbones of the same length, independently from any particular protein. We generated our first test instance by choosing randomly one of the leaf nodes (solutions) on this tree. We then constructed its three-dimensional conformation, and we computed the distances between all its hydrogens. For all the distances d smaller than 5Å, we created an interval $[d-\varepsilon, d+\varepsilon]$ containing the computed distance, and we added it to our set of distances that will be used for pruning. Fig. 6 shows the generated test instance.

The positions of the first three atoms can be obtained by using the known information on the bond lengths and bond angles. The branching starts at level 4, in correspondence with the atom C^1_α (see Table I). At level 5 we have the first duplicated atom, the nitrogen N^1 which already appeared at level 1. Therefore, we have no branching, because the new copy of N^1 can only be placed in the same position of its previous copy.

The first hydrogen in the vertex ordering on which we need to branch appears at level 6. This is the hydrogen H^1_α. Since

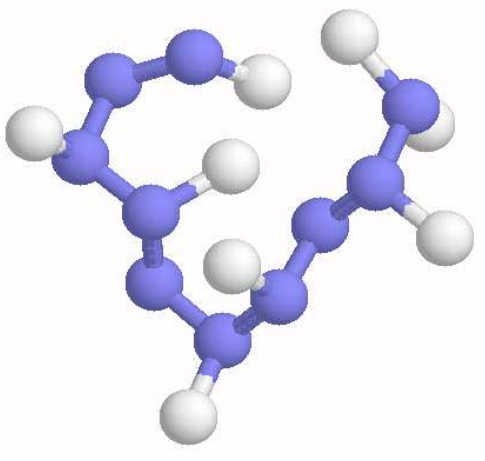

Fig. 6. A generated test instance with 3 amino acids.

the distance between this atom and the previous H^1 is an interval, we need to discretize the interval and take from it a certain number of sample distances, which will be considered as exact. Let us denote by D the number of considered sample distances. As a consequence, $2 \times D$ branches are added at level 6 on the binary tree. At level 7, we find another duplicated atom, and therefore, there is no branching. After this atom, we have a sequence of 3 atoms that are neither duplicated nor hydrogens: depending on the fact that an interval needs to be discretized or not, only two or $2 \times D$ branches are added to the tree.

The first hydrogen of the second amino acid is at level 11. Since the distance between C^1 and H^2 is known a priori, we have only two branches. The other cases are similar to the previous ones.

Table I provides the number of branches on each layer of the tree. We consider here the generated instance in Fig. 6, where $\varepsilon = 0.30$ and $D = 6$. In particular, the last but one column of the table shows the number of branches of the full tree, in which no kinds of prunings are applied. The last column, instead, shows how we can prune by exploiting the distances between hydrogen atoms that have been artificially generated as explained above. It is easy to identify in the table the three different situations that we can have. When an atom is duplicated (see for example the N at level 5), no branches are added to the tree. When the atom is not duplicated and all the distances for the discretization are exact (see for example the C at level 8), we introduce two new branches. Finally, when the atom is not duplicated and an interval needs to be discretized (see for example the H_α at level 6), $24 = 2 \times (2 \times D)$ branches are added to the tree. Without pruning, the tree reaches 9172942848 branches at level 28.

In the last column of Table I we can see the effect of the pruning phase. Every time we consider a hydrogen, there is a good chance to have a distance that regards this hydrogen. This distance (represented by an interval) can be used for

layer	*atom*	*duplicated?*	*w/out pruning*	*with pruning*
1	N	no	1	1
2	H	no	1	1
3	H	no	1	1
4	C_α	no	2	2
5	N	yes	2	2
6	H_α	no	24	**18**
7	C_α	yes	24	18
8	C	no	48	36
9	N	no	576	360
10	C_α	no	1152	720
11	H	no	2304	**10**
12	N	yes	2304	10
13	C_α	yes	2304	10
14	H_α	no	27648	**70**
15	C	no	55296	140
16	C_α	yes	55296	140
17	N	no	663552	1400
18	C	yes	663552	1400
19	C_α	no	1327104	2800
20	H	no	2654208	**4**
21	N	yes	2654208	4
22	C_α	yes	2654208	4
23	H_α	no	31850496	**9**
24	C	no	63700992	18
25	C_α	yes	63700992	18
26	O	no	764411904	52
27	C	yes	764411904	52
28	H	no	9172942848	**10**

TABLE I
THE NUMBER OF BRANCHES, STEP BY STEP, OF THE DISCRETE DOMAIN WITH AND WITHOUT PRUNING.

pruning away all the branches containing infeasible solutions. We prune, for example, at level 11 when considering the hydrogen H of the second amino acid. In the previous layer of the tree, 720 branches are contained. At level 11, two branches are added to the ones of the previous layer. As a consequence, 1440 branches are considered in total, but only 10 branches pass the pruning test. Similarly, at level 14, 20 and 28, the pruning phase allows to drastically reduce the number of branches. The pruned tree has only 10 leaf nodes, which represent the 10 solutions related to our small instance.

Algorithm IV is an extension of the BP algorithm, previously proposed in [9], for considering interval data. It is naturally implied from the discussion above. Our implementation of this extension of the BP algorithm is able to find the 10 solutions for the instance detailed above in less than 1 second of CPU time.

We also performed some experiments by considering larger instances. The procedure which has been employed for generating such instances is exactly the same as before, with the only difference that the ordering for the generic amino acid (see Figure 4) is repeated as many times as needed. Naturally, these instances do not represent well real protein backbones for the general shape they have, but they are still useful for the purposes of the experiments. The experiments (see Table II) showed that the discretization with interval data can also be applied to instances having a larger dimension. In the table,

```
0: branch-and-prune(i, n, d, nbranches)
  if (x_i is a duplicated atom) then
    assign to x_i the same coordinates of its previous copy;
    branch-and-prune(i+1,n,d,nbranches);
  else
    if (d(i-3,i) is exact) then
      b = 2;
    else
      b = nbranches;
    end if
    for (k = 1, b) 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 (x_i^k is feasible) then
        if (i = n) then
          a solution is found;
        else
          branch-and-prune(i+1,n,d,nbranches);
        end if
      else
        the current branch is pruned;
      end if
    end for
  end if
```

n_{aa} is the total number of amino acids, n is the total number (including the repetitions) of considered atoms and $|E|$ is the number of distances which are available (exact distances and intervals). We only require one solution, and therefore #Sol is always 1. Finally, the CPU time increases to almost half an hour when the largest instance (with 1000 amino acids) is considered. We point out that the experiments have been performed on an Intel Core 2 CPU 6400 @ 2.13 GHz with 4GB RAM, running Linux.

V. CONCLUSION

In this paper, we defined an artificial ordering for discretizing MDGPs with interval data. This ordering allowed us to solve two issues which arose while working on other discretization approaches. First, we are now able to consider interval data provided by NMR experiments. Secondly, we are now able to consider the hydrogens of the protein backbones together with the other backbone atoms, which allows to avoid the numerical instabilities of the previously proposed approach [8], based on the solution of a sequence of linear systems.

n_{aa}	n	$\|E\|$	#Sol	time
10	91	716	1	0.1 s
100	901	7556	1	0.6 s
1000	9001	75956	1	27 m

TABLE II
SOME EXPERIMENTS WITH LARGER INSTANCES. ONLY ONE SOLUTION IS REQUIRED.

ACKNOWLEDGMENTS

The authors wish to thank Leandro Martinez for his valuable comments on this work. 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] 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.

[2] I.D. Coope. Reliable computation of the points of intersection of n spheres in $\mathbb{R}^n$. *Australian and New Zealand Industrial and Applied Mathematics Journal*, 42:C461–C477, 2000.

[3] S. Kirkpatrick, C.D.Jr. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[4] C. Lavor, L. Liberti, and N. Maculan. The discretizable molecular distance geometry problem. Technical Report q-bio/0608012, arXiv, 2006.

[5] 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, second edition, 2009.

[6] C. Lavor, A. Mucherino, L. Liberti, and N. Maculan. An artificial backbone of hydrogens for finding the conformation of protein molecules. In *Proceedings of the Computational Structural Bioinformatics Workshop*, pages 152–155, Washington D.C., USA, 2009. IEEE BIBM.

[7] 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*, pages 751–756, Mragowo, Poland, 2009. IEEE.

[8] C. Lavor, A. Mucherino, L. Liberti, and N. Maculan. On the computation of protein backbones by using artificial backbones of hydrogens. *Journal of Global Optimization*, accepted.

[9] 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.

[10] L. Liberti, C. Lavor, A. Mucherino, and N. Maculan. Molecular distance geometry methods: from continuous to discrete. *International Transactions in Operational Research*, to appear.

[11] 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.

[12] A. Mucherino, C. Lavor, L. Liberti, and N. Maculan. On the definition of artificial backbones for the discretizable molecular distance geometry problem. *Mathematica Balkanica*, 23:289–302, 2009.

[13] A. Mucherino, C. Lavor, L. Liberti, and N. Maculan. Strategies for solving distance geometry problems with inexact distances by discrete approaches. In *Proceedings of the conference TOGO10*, pages 93–96, Toulouse, France, 2010.

[14] M. Nilges, A.M. Gronenborn, A.T. Brunger, and G.M. Clore. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Protein Engineering*, 2:27–38, 1988.

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