An Artificial Backbone of Hydrogens for Finding the Conformation of Protein Molecules

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

*Dept. of Applied Mathematics (IMECC-UNICAMP), State University of Campinas, Campinas-SP, Brazil.
clavor@ime.unicamp.br
†LIX, École Polytechnique, Palaiseau, France.
{mucherino,liberti}@lix.polytechnique.fr
‡COPPE, Systems Engineering, Federal University of Rio de Janeiro, Rio de Janeiro-RJ, Brazil.
maculan@cos.ufrj.br

***Abstract*—NMR experiments can provide distances between pairs of hydrogens of a protein molecule. The problem of identifying the coordinates of such hydrogens by exploiting the information on the distances is a Molecular Distance Geometry Problem (MDGP). In a previous work, we defined an artificial backbone of hydrogens related to the protein backbones, where a particular ordering was given to the hydrogens. This ordering allows to formulate the MDGP as a combinatorial optimization problem, to which we refer as the Discretizable MDGP (DMDGP) and that we efficiently solve by an exact algorithm, the Branch and Prune (BP) algorithm. Once the coordinates of the hydrogens have been found, the problem of finding the remaining backbone atoms (N, C_α and C) is another MDGP. In this short paper, we propose a simple method for solving the MDGP related to the backbone atoms N, C_α and C of a protein, where the coordinates of the hydrogens previously found by the BP algorithm are exploited.**

I. Introduction

Nuclear Magnetic Resonance (NMR) experiments are able to detect the distances between pairs of atoms of a protein molecule. Even though molecules can be formed by different kinds of atoms, NMR usually detects distances between hydrogen atoms shorter than 6Å. All these distances can be used to find the conformation of the molecule. This problem is known in the literature as the Molecular Distance Geometry Problem (MDGP) [1], [3].

In its basic form, the MDGP is a constraint satisfaction problem, because a set of constraints on the distances must be satisfied in the possible solutions to the problem. However, the problem is usually faced by global continuous optimization techniques, where a penalty function is defined in order to measure how much a given conformation satisfies the constraints (for a survey on methods for the MDGP, see [6]). The global minima of the penalty function correspond to the conformations in which all the constraints (or the majority of the constraints) are satisfied. One of the most used penalty function is the Largest Distance Error (LDE):

$$LDE(\{x_1, x_2, \ldots, x_n\}) = \frac{1}{m} \sum_{\{i,j\}} \frac{|\,||x_i - x_j|| - d_{ij}\,|}{d_{ij}},$$

where m is the total number of available distances, x_i is the generic atom of the conformation, d_{ij} is the known distance between x_i and x_j, and $||x_i - x_j||$ is the computed distance between x_i and x_j.

We recently proposed the Discretizable MDGP (DMDGP) [5], [7], [9], [11], [12], which consists in a subclass of instances of the MDGP for which a discrete formulation can be supplied. This subclass contains instances that satisfy two particular assumptions:

- for each quadruplet of consecutive atoms, all the relative distances must be known;
- for each triplet of consecutive atoms x_1, x_2 and x_3, the distance between x_1 and x_3 cannot be perfectly equal to sum of the distance between x_1 and x_2 and the distance between x_2 and x_3.

In practice, the second assumption is always satisfied, especially when the distances are considered in the floating-point arithmetic of a computer machine. The first assumption, instead, is harder to be satisfied.

The reformulation of the MDGP as a combinatorial problem allows to reduce the search domain from a continuous to discrete set. However, both the MDGP and the DMDGP are NP-hard problems [5], [13]. The DMDGP has particular symmetry properties that can be exploited in order to find solutions to the problem in a more efficient way [5].

The Branch and Prune (BP) algorithm [9] is an exact algorithm for the DMDGP. Our computational experiences showed that it is very efficient in solving the discrete problem, in both terms of CPU time and quality of the solutions [5], [7], [9]. We also implemented a modified version of the algorithm which is able to handle experimental errors [12]. In its basic form, the BP algorithm is able to deal with exact values d_{ij}

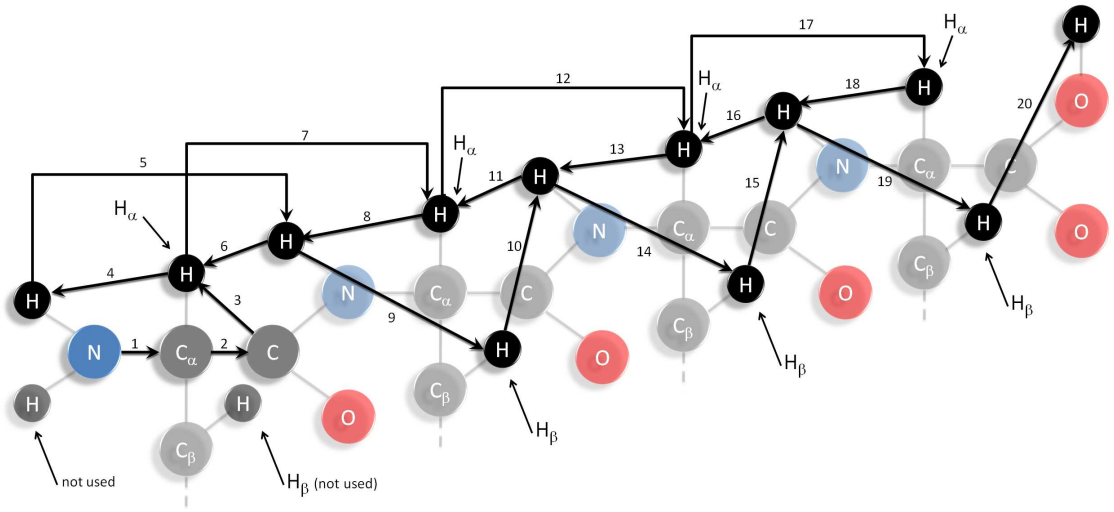

Fig. 1. An artificial backbone of hydrogens related to the protein backbones. Note that some of the hydrogens are considered twice and that the considered ordering is specified by the labels associated to the edges.

for the known distances that form an instance of the problem. Moreover, we are also working on a new version of the BP algorithm that can manage intervals $[l_{ij}, u_{ij}]$, where the distances d_{ij} are contained, instead of exact distances [10].

We proposed in [8] a special ordering for the hydrogens related to the backbones of protein molecules for which the two assumptions for the DMDGP are satisfied. We named this sequence of atoms *artificial backbone of hydrogens*. Since the two necessary assumptions are satisfied, the problem of finding the coordinates of all the atoms of this artificial backbone is a DMDGP. As a consequence, we can solve this problem by applying the BP algorithm.

In this short paper, we present a method for constructing the real backbone of a protein conformation from the coordinates of its hydrogens. Let us suppose that NMR found the relative distances between the hydrogens related to the backbone of a certain protein. If the hydrogens are sorted by following the ordering defined by the artificial backbone, the problem of identifying the coordinates of these hydrogens is a DMDGP. Once such coordinates have been obtained by applying the BP algorithm, the remaining backbone atoms, and in particular the sequence of atoms $\mathrm{N} - \mathrm{C}_\alpha - \mathrm{C}$, can be obtained by solving another MDGP. This MDGP is easier to solve, because assumptions stronger than the ones needed for the DMDGP are satisfied. We will show how efficiently solve this MDGP.

The rest of the paper is organized as follows. In Section II we briefly introduce the artificial backbone of hydrogens proposed in [8]. Then, in Section III, we show a simple method for constructing the real backbone of a protein starting from the coordinates of its hydrogens. Preliminary computational experiments are shown in Section IV and conclusions are given in Section V.

II. An artificial backbone of hydrogens

The artificial backbone presented in [8] considers only the hydrogen atoms related to the real backbone of a protein. There are 4 hydrogens that are common to all the amino acids. 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) [14]. Therefore, only two hydrogens per amino acid are contained in the backbone of a protein: a hydrogen H bound to N, and a hydrogen H_α bound to C_α.

The two hydrogens H and H_α are used for defining the artificial backbone. Moreover, another hydrogen per amino acid is borrowed from the amino acid side chain. Each amino acid has a different side chain: 19 of the 20 amino acids involved in the protein synthesis have a carbon atom C_β in their side chains which is bound to the carbon atom C_α. At least one hydrogen is bound to the carbon C_β, and one of them is considered 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*, the third considered hydrogen is the only one that forms the side chain of this amino acid. We refer to this hydrogen with the symbol H_β.

The artificial backbone of hydrogens is shown in Figure 1. The same hydrogen can be considered twice in the sequence in order to reduce the relative distances between pairs of hydrogens. Moreover, note that the first three atoms of the artificial backbone are not hydrogen atoms. We added them because they define a common coordinate system for the real backbone of the protein and the artificial backbone of hydrogens. The distances related to these three atoms, needed for formulating the problem as a DMDGP, do not need to be detected by NMR, because they are known a priori.

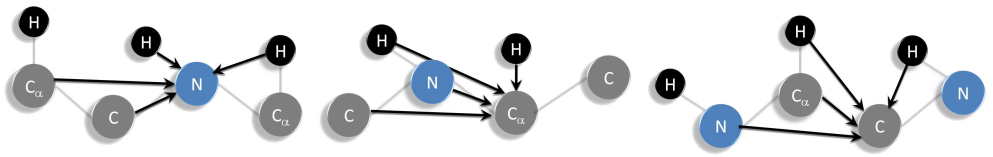

Fig. 2. The atoms and the distances used in the three linear systems used to determine the protein backbone.

III. Computing the protein backbone from the artificial one

Let us suppose that the coordinates of the hydrogens of an artificial backbone have been obtained by BP. The problem of finding the coordinates of the backbone atoms N, C_α and C is a MDGP. However, since the coordinates of the hydrogens are already known, and some distances between hydrogens and the other backbone atoms are known a priori, this MDGP satisfies assumptions that are stronger than the ones of the DMDGP. As a consequence, the MDGP related to the backbone atoms $N - C_\alpha - C$ can be solved in linear time, by following the method presented in [2], [15]. Let us suppose that we need to find the coordinates of the backbone atom a and that the distances between a and the 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.

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

$$\begin{cases} ||a-b_1|| &= d_{a,b_1} \\ ||a-b_2|| &= d_{a,b_2} \\ ||a-b_3|| &= d_{a,b_3} \\ ||a-b_4|| &= d_{a,b_4}. \end{cases} \quad (1)$$

This is a quadratic system of 4 equations in 3 variables. However, as shown in [2], [15], if the system of linear equations

$$Ax = b, \quad (2)$$

where

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

$$x = a,$$

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

is solved, its solution is also solution for the quadratic system (1). Thus, the MDGP related to the protein backbone can be solved by solving a sequence of 3×3 linear systems.

For each atom N, C_α and C of the protein backbone, there are 4 atoms b_i that need to be considered in the linear system. 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_α and C of the previous amino acid, the hydrogen H bound to N and the hydrogen H_α bound to the following C_α (see Figure 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 linear system (2) allows to identify the coordinates of N. Similar observations can be made for the other two systems, related to the backbone atoms C_α and C. See Figure 2 to find out which atoms and distances can be considered.

IV. Preliminary computational experiences

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 very similar to the one proposed in [4]. 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 II. Only distances smaller than 6Å are considered. We randomly generated a set of instances having a different number of amino acids. For each amino acid, 3 hydrogen atoms are defined, and 5 in total are included in the instance. 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 I. In the table, n is the number of atoms included in the instance. It is always a number which is

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

TABLE I
RESULTS OBTAINED BY THE BP ALGORITHM ON A SET OF RANDOMLY GENERATED ARTIFICIAL BACKBONES.

divisible by 5, because each amino acid of the considered artificial backbone contains exactly 5 hydrogens (two of them are considered twice). m is the number of known distances, and the LDE function (modified in order to avoid divisions by zero, see [8] for more details) is used for evaluating the quality of the solution. 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.

For each found solution, we applied the method discussed in Section III and we recontructed the real backbones corresponding to the artificial ones. The software procedure we used just solves a sequence of linear systems. These experiments show that all the atoms of the protein backbones can be computed starting from the information obtained from NMR experiments, that mainly regard hydrogen atoms.

V. CONCLUSIONS

Data from NMR experiments can be used for finding the conformations of molecules and, in particular, of proteins. Such data mostly regard the hydrogens contained in the molecule, and therefore the first problem that one can solve is related to the subset of hydrogens of the molecule. We showed in previous works that this can be efficiently done by organizing the hydrogens on an artificial backbone and by reformulating the problem as combinatorial. Moreover, we showed in this work that the other atoms bound to the hydrogens can be successively identified by solving a sequence of linear systems.

In this short paper, our focus was on the protein backbones. This idea can also be extended to entire protein conformations. To this aim, the artificial backbone of hydrogens needs to be extended in order to consider all the hydrogens in the protein, and the method presented in this paper needs to be extended for considering the side chains of the amino acids. We are also working on a formal definition of the artificial backbone and on proofs showing that it belongs to the DMDGP subclass. We plan to present these results in future publications.

ACKNOWLEDGMENTS

The authors would like to thank the Brazilian research agencies FAPESP and CNPq, the French research agency CNRS and École Polytechnique, for financial support.

REFERENCES

[1] G.M. Crippen and T.F. Havel, *Distance Geometry and Molecular Conformation*, John Wiley & Sons, New York, 1988.
[2] 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.
[3] T.F. Havel, *Distance Geometry*, D.M. Grant and R.K. Harris (Eds.), Encyclopedia of Nuclear Magnetic Resonance, Wiley, New York, 1701-1710, 1995.
[4] 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.
[5] C. Lavor, L. Liberti, and N. Maculan, *Discretizable Molecular Distance Geometry Problem*, Tech. Rep. q-bio.BM/0608012, arXiv, 2006.
[6] 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.
[7] 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.
[8] 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, October 2009.
[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), 1–17, 2008.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] T. Schlick, *Molecular Modelling and Simulation: an Interdisciplinary Guide*, Springer, New York, 2002.
[15] 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.