The Discretizable Molecular Distance Geometry Problem : from Ideal to Real Instances

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

[1] CERFACS, Toulouse, France, mucherino@cerfacs.fr
[2] IMECC, UNICAMP, Campinas, Brazil, clavor@ime.unicamp.br
[3] LIX, École Polytechnique, Palaiseau, France, liberti@lix.polytechnique.fr

Mots-clés : *distance geometry, discretization, protein conformation, Branch & Prune, NMR data.*

The Molecular Distance Geometry Problem (MDGP) is the problem of finding the conformation of a molecule (i.e. the coordinates in the three-dimensional space of all its atoms) by using known relative distances between pairs of its atoms. The information regarding the relative distances can be obtained by experimental techniques such as Nuclear Magnetic Resonance (NMR) spectroscopy. The following steps are usually performed in order to solve an MDGP : *(i)* perform NMR experiments ; *(ii)* create an instance of the MDGP by exploiting the results obtained through the NMR experiments ; *(iii)* solve the instance by employing a suitable method. Many methods and algorithms for the solution of MDGPs have been proposed, and the majority of them are based on a continuous formulation of the problem [6].

In the last five years, we have been working on a combinatorial reformulation for the MDGP, to which we refer as Discretizable MDGP (DMDGP) [1]. When some particular assumptions are satisfied, which are strongly based on the ordering given to the atoms of the molecule, we are able to reduce the search domain (where the possible conformations for the molecule can be found) to a binary tree. Even though the MDGP and the DMDGP are both NP-hard [1], the discretization allows us to employ a Branch & Prune (BP) algorithm, that is very efficient in finding the solution (and, in general, the complete set of solutions) to a DMDGP [5]. Distances contained into an instance of the DMDGP can be divided in two subsets : a subset containing all the distances that are necessary for the discretization, which are used for constructing the binary tree, and the other one containing other distances which can be exploited for pruning away infeasible branches of the tree. The basic pruning test in the BP algorithm, to which we refer as Direct Distance Feasibility (DDF), verifies if known and computed distances match.

First computational experiments on *ideal* instances showed the efficiency and reliability of the BP algorithm [1, 2, 5]. However, such instances were *ideal* in the sense that they had some properties which allowed us to discretize the MDGP quite easily. For example, we were supposing that all the available distances between pairs of atoms were exact. Moreover, we were not making any distinction among the different kinds of atoms that can compose a molecule. In general, only noisy distances between pairs of hydrogen atoms can instead be found through NMR experiments.

The step from *ideal* instances to *real* instances (that actually contain data from NMR) has not been trivial. The discretization process is based on the computation of the intersection of three spheres in the three-dimensional space. If the distances are represented by intervals, the three spheres become three spherical shells, whose intersection is a geometrical object whose shape is, to the best of our knowledge, unknown. Moreover, since distances between hydrogens are mainly provided by NMR, many distances needed for the discretization may not be available because regarding pairs of atoms which are not hydrogens.

Starting from the ideal instances, we began to consider instances closer and closer to the real ones. In [7], for example, we managed for the first time interval data, but information on all kinds of atoms were available in our instances. In [4], we considered instances where only hydrogen atoms are included, and we divided the MDGP in two subproblems : the one of finding the coordinates of all the hydrogen atoms (reformulated as a DMDGP), and the one of finding the coordinates of the other atoms, that can be solved in polynomial time. In this study, however, all distances were supposed to be exact.

Recently, we have been able to make the last step and get very close to real instances [3]. We found a hand-craft ordering for the atoms of the *protein backbones* which allows to perform the discretization when only distances between hydrogens are available and when they are all represented by intervals. Known bond lengths and bond angles for the atoms of the protein can be used for deriving a subset of distances and intervals that can be exploited for performing the discretization. Our hand-craft ordering is designed in such a way that only these distances and intervals are used for the discretization of the problem. Moreover, only one interval per time is involved, and therefore the discretization process consists in the intersection of two spheres with one spherical shell. This intersection defines a curve in the three-dimensional space, whose shape is unknown. Therefore, we rather discretize the involved interval and take some sample distances from it, in order to reduce our problem to several (depending on the number of sample distances) intersections among spheres. Distances between hydrogens obtained by NMR are only used for pruning purposes : the DDF pruning test can be trivially adapted to interval data.

Instances of this kind are very similar to instances of the problem which biologists and chemists need to solve for finding the conformation of a molecule from NMR data. Our preliminary experiments with artificially generated instances are very promising [3]. Next step is to use real data obtained by NMR experiments and try to reproduce the conformation of well-known molecules by using our discrete approach. Moreover, since the BP algorithm is potentially able to find all possible solutions to the problem (differently from algorithms based on continuous approaches and/or heuristics), we might be able to identify some possible conformations for these molecules that were not previously detected.

Références

[1] C. Lavor, L. Liberti, and N. Maculan, *Discretizable Molecular Distance Geometry Problem*, Tech. Rep. q-bio.BM/0608012, arXiv, 2006.

[2] 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.

[3] C. Lavor, L. Liberti, A. Mucherino, *On the Solution of Molecular Distance Geometry Problems with Interval Data*, IEEE conference proceedings, International Conference on Bioinformatics & Biomedicine (BIBM10), Hong Kong, 77–82, 2010.

[4] C. Lavor, A. Mucherino, L. Liberti, N. Maculan, *On the Computation of Protein Backbones by using Artificial Backbones of Hydrogens*, to appear in Journal of Global Optimization, 2010. Published online on July 24, 2010.

[5] 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.

[6] L. Liberti, C. Lavor, A. Mucherino, N. Maculan, *Molecular Distance Geometry Methods : from Continuous to Discrete*, International Transactions in Operational Research **18**(1), 33–51, 2011.

[7] A. Mucherino, C. Lavor, *The Branch and Prune Algorithm for the Molecular Distance Geometry Problem with Inexact Distances*, Proceedings of World Academy of Science, Engineering and Technology (WASET), International Conference on Bioinformatics and Biomedicine (ICBB09), Venice, Italy, 349–353, 2009.