On a Discretizable Subclass of Instances of the Molecular Distance Geometry Problem

Carlile Lavor
Dept. of Applied Mathematics,
State University of Campinas,
Campinas-SP, Brazil
clavor@ime.unicamp.br

Leo Liberti
Antonio Mucherino
LIX, École Polytechnique,
Palaiseau, France
{liberti,mucherino}@
lix.polytechnique.fr

Nelson Maculan
COPPE
Systems Engineering,
Federal University of
Rio de Janeiro,
Rio de Janeiro, Brazil
maculan@cos.ufrj.br

ABSTRACT

The molecular distance geometry problem can be formulated as the problem of finding an immersion in $\mathbb{R}^3$ of a given undirected, nonnegatively weighted graph G. In this paper, we discuss a set of graphs G for which the problem may also be formulated as a combinatorial search in discrete space. This is theoretically interesting as an example of "combinatorialization" of a continuous nonlinear problem. It is also algorithmically interesting because the natural combinatorial solution algorithm performs much better than a global optimization approach on the continuous formulation. We present a Branch and Prune algorithm which can be used for obtaining a set of positions of the atoms of protein conformations when only some of the distances between the atoms are known.

Categories and Subject Descriptors

J.3 [**Life and medical sciences**]: Biology and genetics; G.2.1 [**Combinatorics**]: Combinatorial algorithms; G.2.2 [**Graph Theory**]: Graph algorithms; G.1.6 [**Optimization**]: Global optimization

General Terms

algorithms, performance, experimentation

Keywords

distance geometry, protein molecules, protein backbone, undirected weighted graph, combinatorialization.

1. INTRODUCTION

The Molecular Distance Geometry Problem (MDGP) is the problem of finding the positions of the atoms of a molecular conformation when only some of the distances between such atoms are known. In literature, different approaches for solving the MDGP have been proposed. The reader is referred to [3] for a survey. The MDGP is usually formulated as a continuous nonconvex optimization problem, where the objective function evaluates the differences between the known distances $d(u,v)$ and the distances $||x(u)-x(v)||$ implied by a possible conformation $X = \{x_1, x_2, \dots, x_n\}$. One of the choices for the objective function is

$$g(X) = \frac{1}{|V|} \sum_{\{u,v\} \in E} \frac{||x(u) - x(v)|| - d(u,v)}{d(u,v)}. \qquad (1)$$

where u and v refer to two different atoms, and the set E contains all the couples of atoms whose distance is known. Note that X is solution of the MDGP if and only if $g(X) = 0$.

We consider the following formulation of the MDGP problem. Let $G = (V, E, d)$ be a weighted undirected graph. The vertices in V represent the atoms in a conformation, the links in E represent the couples of atoms whose relative distance is known, and the weights d are the known distances. The MDGP can be seen as the problem of finding an immersion $x : G \to \mathbb{R}^3$ such that $||x(u) - x(v)|| = d(u,v)$ for each $\{u,v\} \in E$.

Instead of using a continuous formulation of the problem, we introduce the Discretizable Molecular Distance Geometry Problem (DMDGP), which consists of a certain subset of MDGP instances for which a discrete formulation can be supplied. Given a generic instance, let $G = (V, E, d)$ be the associated weighted undirected graph. The instances we take into account satisfy two assumptions: 1) E contains all cliques on quadruplets of consecutive vertices; 2) the strict triangular inequality $d(v_{i-1}, v_{i+1}) < d(v_{i-1}, v_i) + d(v_i, v_{i+1})$ holds for $i = 2, \dots, n-1$. If these two assumptions are satisfied, each atom $v \in V$ can be located at most in two different positions. In this way, we do not have to deal with continuous variables, but rather with binary variables. This allows to build a *tree* of possible choices where the solutions of the DMDGP can be searched.

We solve instances of the DMDGP by a Branch and Prune (BP) algorithm. We consider instances providing a set of distances between the atoms of the protein backbones. Indeed, the structure of the protein backbones makes it possible to formulate the problem as a DMDGP, because most of these instances satisfy our assumptions. It should be noted straight away that we refer here to the distance geometry problem as a precisely formalized decision prob-

SAC'09 March 8-12, 2009, Honolulu, Hawaii,U.S.A.
Copyright 2009 ACM 978-1-60558-166-8/09/03 ...$5.00.

Table 1: Some computational experiences.

test	ϵ	*obj*	#P1	*time*	#P1	#P2	*time*
L100(1)	10^{-3}	$\tilde{}10^{-10}$	51	0s	51	0	0s
L100(2)	10^{-3}	$\tilde{}10^{-9}$	815010	1.7s	52673	4131	72.5s
L100(3)	10^{-3}	$\tilde{}10^{-9}$	53	0s	53	0	0.1s
L200(1)	10^{-3}	$\tilde{}10^{-8}$	918	0s	395	36	0.8s
L200(2)	10^{-3}	$\tilde{}10^{-8}$	394	0s	117	7	3.4s
L200(3)	10^{-3}	$\tilde{}10^{-9}$	4732	0s	1791	48	2.1s

lem, and not as a practical chemical problem. However, our computational experiments have been successfully carried out on instances generated from protein conformations.

In Section 2 some details about the BP algorithm are provided, while some computational results are shown in Section 3. Section 4 concludes this short paper. Refer to [2, 3, 4] for more details on the DMDGP and the BP algorithm.

2. THE BRANCH AND PRUNE ALGORITHM

Given an instance of the DMDGP, the BP algorithm searches for its solutions as follows. As previously observed, there are only two possible choices for each atomic position, x_i and x'_i. These positions can be, however, either feasible or infeasible with respect to the distances d. Three situations are possible: 1) both x_i and x'_i are feasible: we add two branches to the tree of the possible choices, and both of them are explored in a depth-first fashion; 2) only one of the positions is feasible: we continue the search on the branch defined by this atom position; 3) both x_i and x'_i are infeasible: the current branch is pruned and the search is backtracked.

We implement two pruning tests for checking the feasibility of the atom positions. The first one (P1) checks whether the known distances d and the distances implied by the chosen positions are the same or not. For doing so, the inequality $|||x_v - x_u|| - d_{v,u}| < \varepsilon$ is controlled for all the couples $(u, v) \in E$, where ε is a positive tolerance.

The second pruning test (P2) is based on the Dijkstra shortest-path searches on Euclidean graphs. Consider atoms h, i, k with $h < i < k$ such that the distance d_{hk} is known. Suppose that the algorithm already placed the atom h, and that the feasibility of the atom i needs to be verified. Let $D(i, k)$ be an upper bound to the distance $||x_i - x_k||$ for all possible immersions $x : G \to \mathbb{R}^3$ which are feasible solutions. If the inequality $D(i, k) < ||x_h - x_i|| - d_{hk}$ holds, then the search node for the atomic position x_i can be pruned [5]. We use as upper bound $D(i, k)$ the shortest path between the node i and the node k.

3. COMPUTATIONAL EXPERIENCES

A software procedure has been developed in C programming language which implements the BP algorithm. Computational experiences have been performed on different sets of instances. The instances generated as explained in [1] have been used for comparing the performances of the algorithm when the pruning test P1 is used alone or coupled with the pruning test P2. We always used P1 because it is a very natural way to prune atom positions and it is also easy to implement. P2 is more complex.

Table 1 shows some computational results. In these experiments, the BP algorithm is stopped as soon as it finds the first solution. Note that the same instance might have more than one solution. The same experiments have been performed by exploiting only P1, or P1 and P2 in cooperation. In the table, #P1 indicates the number of times an atom position is pruned by P1, and #P2 has the same meaning regarding to P2. The tolerance ϵ in P1 is always set to 10^{-3}, and *obj* is the value of the objective function (1) in the found solution. The experiments have been carried out on a Intel Core 2 CPU @ 2.13GHz with 4GB RAM, running Linux.

Table 1 shows that the *obj* values in the solutions range from 10^{-8} to 10^{-10}. These values do not change at all if only P1 is applied or the two pruning tests are used together. In general, the number of atoms pruned by P1 alone is always greater or equal to the number of atoms pruned by P1 and P2 in collaboration. This indicates that the pruning test P2 is more efficient, because it is able to identify infeasible atoms better and prune them earlier on the search tree. This efficiency of the pruning test P2 is unfortunately not reflected on the computational time. Indeed, even though less atoms are pruned in total when both P1 and P2 are used, the computational time increases, since shortest paths need to be computed.

We applied the BP algorithm for solving instances generated from real protein conformations. The algorithm is able to find solutions with a good accuracy. The only issue arising when real instances are considered is that the known distances may not be so accurate, because of possible experimental errors.

4. CONCLUSIONS

In this paper we presented a new discrete formulation for a particular subclass of the MDGP. We proposed the BP algorithm for solving it and provided computational experiences. We also showed the effectiveness and efficiency of two possible pruning tests. Future research will be devoted to techniques for managing the experimental errors contained into real data.

Acknowledgments

The authors would like to thank CNPq and FAPESP for their financial support.

5. REFERENCES

[1] C. Lavor, *On Generating Instances for the Molecular Distance Geometry Problem*, In: L. Liberti and N. Maculan (Eds.), Global Optimization: from Theory to Implementation, Springer, New York, 405–414, 2006.

[2] C. Lavor, L. Liberti, and N. Maculan, *Computational Experience with the Molecular Distance Geometry Problem*, In: Global Optimization: Scientific and Engineering Case Studies, J. Pintér (Ed.), 213–225. Springer, Berlin, 2006.

[3] C. Lavor, L. Liberti, and N. Maculan, *An Overview of Distinct Approaches for the Molecular Distance Geometry Problem*, In: Encyclopedia of Optimization, C. Floudas and P. Pardalos (Eds.), 2^{nd} edition, Springer, New York, to appear.

[4] 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.

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