Finding low-energy homopolymer conformations by a discrete approach

C. Lavor,[1] A. Mucherino,[2] L. Liberti,[3] N. Maculan[4]

[1] *IMECC-UNICAMP, Campinas-SP, Brazil,* clavor@ime.unicamp.br

[2] *IRISA, University of Rennes 1, Rennes, France,* antonio.mucherino@irisa.fr

[3] *LIX, École Poleytechnique, Palaiseau, France,* liberti@lix.polytechnique.fr

[4] *COPPE, Federal University of Rio de Janeiro, Rio de Janeiro-RJ, Brazil,* maculan@cos.ufrj.br

Abstract Finding energetically stable conformations of proteins is one of the most interesting and difficult problems in biology. Given the chemical composition of a molecule, its three-dimensional conformation is of interest because it is directly related to the function it is able to perform. Since some years, we have been working on an easier problem, where the identification of protein conformations is aided by information obtained by experiments of Nuclear Magnetic Resonance (NMR). Under certain assumptions, we are able to discretize the problem and use an efficient Branch & Prune (BP) algorithm for its solution. In this work, we move towards the more difficult situation in which the information given by NMR is not available. The original BP does not work in this case, because there is no information for performing its pruning phase, that is the strong point of the algorithm. In this paper, we study and present a new energy-based pruning device to be added to the BP algorithm. First computational experiences on a set of small homopolymers show that this approach is promising for the identification of low-energy conformations of molecules.

Keywords: homopolymer conformation, energy minimization, discretization, branch-and-prune

1. Introduction

Proteins are composed by smaller molecules called amino acids. They are able to perform several important functions in bodies of living beings, and the functions they perform are strongly dependent on their three-dimensional conformations. While modern technologies are currently able to easily identify the sequence of amino acids for a given protein, the identification of the corresponding three-dimensional conformation is still a real challenge.

There are different approaches to this problem, which depend on the information that are actually available and that can be efficiently used for its solution. Experimental techniques, such as the Nuclear Magnetic Resonance (NMR), are able to provide a subset of lower and upper bounds on some inter-atomic distances, and this information can be exploited for identifying possible three-dimensional conformations for a given molecule. However, if the only knowledge about the molecule is given by its chemical composition, the only way to find an approximation of its conformation is by identifying the most stable conformation from an energetic point of view.

Since years, we are working on the problem of identifying the conformation of a molecule by exploiting information on the distances that NMR experiments are able to provide (see [2–4, 6, 7]). This problem is known in the scientific literature as the Molecular Distance Geometry Problem (MDGP) [1], and it is an NP-hard problem. In this context, our main contribution is given by a discretization process that we can apply for reducing the search space from a continuous domain to a discrete one. Even if this transformation does not decrease the complexity of the problem (which is still NP-hard), it allowed us to conceive an efficient Branch & Prune (BP) algorithm for the solution of MDGPs that can be discretized. We named this class of problems Discretizable MDGP (DMDGP). An instance of the MDGP belongs to the DMDGP class if and only if some assumptions are satisfied [2].

The discretization is performed by intersecting three spheres centered in three consecutive atoms and having three known distances as radii. This intersection, with probability 1, gives two points only, which correspond to the possible coordinates for the successive atom in the sequence [2]. This allows to generate a tree where each layer contains the potential coordinates for the same atom. As a consequence, a solution to the DMDGP can be searched by identifying a path from the root node (representing the first atom) to one of the feasible leaf nodes (representing the last atom) of the tree. Each branch included in the solution path corresponds to a certain torsion angles ω (defined by 4 consecutive atoms of the molecule). The discretization is possible when the reference distances (the radii of the spheres) are exact, as well as when one of the three reference distances is represented by an interval [4].

Even for small-sized molecules, the complete construction of the tree can be too computationally expensive. However, distances (that NMR can provide and that are not exploited in the construction of the tree) can be used in the BP algorithm for pruning a part of the branches where there are no feasible solutions. The pruning phase in BP allows therefore to focus the search on branches of the tree where there can be solutions. In protein instances of the DMDGP, distances used for pruning are available and they allow to prune large parts of the tree, so that the problem can be solved very efficiently by the BP algorithm [7].

In a recent work [4], we proposed a special ordering for the atoms forming protein backbones: all instances in which the atoms are sorted accordingly to this special ordering belong to the DMDGP class. Moreover, all the distances that are necessary for applying the discretization process do not have to be computed by NMR experiments, but they can rather be obtained by simple observations on the chemical composition of protein backbones. Many distances are related to bonded atoms and other ones are related to atoms bonded to a common atom.

It is important to remark that, if this special ordering is employed, the discretization process is completely independent on the DMDGP instance at hand. Actually, any instance related to any protein backbone of a given size has the same tree as a search domain. NMR data can be used for selecting the branches of the tree that are compatible with the distances.

In this paper, we consider for the first time the possibility of replacing distance-based pruning devices with energy-based ones. Since our search trees are independent on NMR, these experiments are not necessary for performing the discretization. We can suppose therefore that the only available information about the molecule concerns its chemical composition. At this point, the only possibility for discarding the infeasible branches of BP trees is through energy-based criteria.

This work opens the doors to more difficult problems in biology for which we could supply a suitable discretization. We will present in Section 2 a new pruning device based on energetic criteria, and we will show some preliminary computational experiments in Section 3.

2. New pruning devices for BP

In the BP algorithm, the idea is to generate the possible atomic coordinates for each atom of the molecule, and to verify their feasibility right away after their generation [6]. In the original BP, the pruning phase is performed by exploiting information on a subset of distances that are not considered in the discretization process (supposed to be obtained through NMR experiments). In this work, we suppose instead that all distances necessary for the discretization are obtained by observations on the chemical structure of the molecule at hand, and that there is no information on other additional distances. Therefore, we need to consider the internal energy of the molecule and the pruning phase needs to be based on this energy.

An accurate description of all interactions among the atoms in a molecule can be very complex. There are however approximations of this potential energy that take into consideration the most important interactions. We will consider the same expression used in [8], given by:

$$E_n = E_{bond} + E_{angle} + E_{torsion} + E_{LJ}, \tag{1}$$

where

$$E_{bond} = \frac{1}{2}\sum_i k_d(d_i - d_0)^2, \qquad E_{angle} = \frac{1}{2}\sum_i k_\theta(\theta_i - \theta_0)^2,$$

$$E_{torsion} = \frac{V_{\bar{n}}}{2}\sum_i \left[1 + \cos(\bar{n}\omega_i + \delta)\right], \qquad E_{LJ} = \sum_{i,j} \varepsilon_{i,j}\left[\left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{12} - 2H_{i,j}\left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{6}\right].$$

The term E_{bond} considers the energy given by the interaction between two bonded atoms. Depending on the kind of atoms, indeed, there is a typical inter-atomic distance for the two atoms. Any modification on this distance makes the energetic term positive. The term models therefore the repulsive force between the two atoms if they get too close to each other, as well as the attractive force between them in case they get too far. The parameter k_d is the “bond stretching” constant, and d_0 is the preferred value for the inter-atomic distance.

The energetic term E_{angle} is used similarly to model the local interactions among three atoms X, Y and Z such that atom X is bonded to Y, and Y is bonded to Z. Depending on the kind of atoms, there is a typical local conformation for the three atoms that correspond to a typical angle between the segment XY and the segment YZ. The term E_{angle} ensures that the value of this angle is close to the typical one by penalizing any modifications (in both directions). The parameter k_θ is the “angle bending” constant, and θ_0 is the preferred value for the bond angle.

The third term $E_{torsion}$ allows to define a certain subset of preferred torsion angles ω for the considered conformations. For example, if $\bar{n} = 3$ and $\delta = 0$, then the preferred torsion angles are 60°, 180° and 300°. The energetic term gives a penalty to all other torsion angle values, and the penalty is proportional to the difference between the selected torsion angle and the closest among the preferred ones. $V_{\bar{n}}$ is a “torsional” constant, which depends on the choice of $\bar{n}$ [8].

Finally, the last term is the Lennard Jones potential [5]. $\varepsilon_{i,j}$ and $\sigma_{i,j}$ are two parameters that can be defined by the relationships between the pairs of atoms (or agglomerate of atoms) which are interacting. The parameter $H_{i,j}$ is related to the hydrophobicity and hydrophilicity of the interacting atoms. We will suppose that $H_{i,j}$ is always equal to 1.

Pruning devices have the difficult role of identifying the atomic positions from which infeasible branches start. During the execution of BP, every time a leaf node is reached, the full set of coordinates for a conformation is available, and hence the energy E_n for this conformation can be computed. Let us suppose that $\hat{E}_n$ is the lowest energy found so far. Our idea is to verify in advance whether new branches of the tree can actually contain conformations with an energy that can be smaller than $\hat{E}_n$. This is done by computing a lower bound on the energy concerning all the conformations belonging to a common branch.

The terms E_{bond}, E_{angle} and $E_{torsion}$ are always positive, and hence the lower bound for their values can be 0. E_{LJ} can be negative, but, depending on the range in which the inter-atomic distances can vary, we can compute an accurate lower bound for the actual value. Let us suppose that we are executing BP and that the current layer is the k^{th}, where we have a partial energy value $E_{n(\leq k)}$ (computed by using the available coordinates) and a lower bound $L_{(>k)}$ on the energy $E_{n(>k)}$. If $E_{n(\leq k)} + L_{(>k)} > \hat{E}_n$, then there is no hope to identify a conformation with an energy smaller than $\hat{E}_n$ by exploring the current branch of the tree. This branch can therefore be pruned. The same strategy can be applied for electrostatic potentials

n	4	5	6	7	8	9	10	11	12
energy (kcal)	-0.12	-0.26	-0.42	-0.64	-0.96	-1.45	-1.99	-2.51	-3.27
time (no devices)	0.0s	0.0s	0.0s	0.6s	19.5s	9min 32s	4h 35min	> 12h	> 12h
time (energy-based device)	0.0s	0.0s	0.0s	0.3s	5.6s	1min 2s	9min 1s	1h 4min	10h 11min

Table 1. Experiments with homopolymers having size n ranging from 4 to 12.

3. Preliminary computational experiments

We present in this section some preliminary experiments. All codes were written in C programming language and all the experiments were carried out on an Intel(R) Xeon(TM) CPU 3.40GHz with 4GB RAM, running Linux. The codes have been compiled by the GNU C compiler v.4.1.1.

We consider homopolymers, that consist of strings of bonded atoms having the same chemical properties. Bond lengths and bond angles are considered fixed, so that the first two terms of the potential energy (1) disappear. All bond lengths are fixed to the preferred value $d_0 = 1.526$Åand all bond angles are fixed to $\theta_0 = 109^{\circ}.47$. Moreover, in the term $E_{torsion}$, $\bar{n} = 3$, $\delta = 0$ and $V_{\bar{n}} = 1.3$. Finally, in the Lennard Jones term, $\varepsilon_{i,j} = 0.181$ and $\sigma_{i,j} = 3.3$.

Table 1 shows some experiments with homopolymers having size n ranging from 4 to 12 (in [8], they consider a little larger instances, but use a much more stronger computer system).

When BP is executed without pruning devices, the computational cost is naturally very expensive. When the energy-based pruning device is instead employed, the same solutions can be identified in a shorter amount of time. For example, the same conformation with $n = 10$ and energy $E_n = -1.99$ kcal can be identified in 9 minutes when our new pruning device is employed, and in about 4 hours and half otherwise. Therefore, the proposed pruning device is actually able to identify and prune the branches of the tree where there cannot be conformations with a lower energy. Future works will be devoted to the development of other pruning devices and ad-hoc strategies for speeding up the search.

Acknowledgments

This work is partially supported by the ANR project “Bip:Bip”. The authors would also like to thank FAPESP, CNPq (Brazil), CNRS and École Polytechnique (France) for financial support.

References

[1] G. Crippen, T. Havel, Distance Geometry and Molecular Conformation, Wiley, New York, 1988.

[2] C. Lavor, L. Liberti, N. Maculan, A. Mucherino, The Discretizable Molecular Distance Geometry Problem, to appear in Computational Optimization and Applications, 2012.

[3] C. Lavor, L. Liberti, N. Maculan, A. Mucherino, Recent Advances on the Discretizable Molecular Distance Geometry Problem, European Journal of Operational Research 219, 698–706, 2012.

[4] C. Lavor, L. Liberti, A. Mucherino, The interval Branch-and-Prune Algorithm for the Discretizable Molecular Distance Geometry Problem with Inexact Distances, to appear in Journal of Global Optimization, 2012.

[5] J.E. Lennard-Jones, Cohesion, Proceedings of the Physical Society 43, 461–482, 1931.

[6] 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.

[7] L. Liberti, B. Masson, C. Lavor, A. Mucherino, Branch-and-Prune Trees with Bounded Width, Proceedings of the 10^{th} Cologne-Twente Workshop on Graphs and Combinatorial Optimization (CTW11), 189–193, 2011.

[8] A.T. Phillips, J.B. Rosen, V.H. Walke, Molecular Structure Determination by Convex Underestimation of Local Energy Minima, In P.M. Pardalos, D. Shalloway, and G. Xue (Eds.), Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding, American Mathematical Society 23, 181–198, 1996.