Collecting Data for Generating Distance Geometry Graphs for Protein Structure Determination

S.B. Hengeveld[1], T. Malliavin[2], L. Liberti[3], A. Mucherino[1]

[1] IRISA, Université de Rennes 1, Rennes, France
{simon.hengeveld,antonio.mucherino}@irisa.fr
[2] LPCT, Université de Lorraine, France
therese.malliavin@univ-lorraine.fr
[3] CNRS LIX, École Polytechnique, Palaiseau, France
liberti@lix.polytechnique.fr

Mots-clés : *Distance Geometry, Protein Structure Determination, Force Fields, NMR.*

Given a simple weighted undirected graph $G = (V, E, d)$ and a positive integer K, the Distance Geometry Problem (DGP) asks whether a realization $x : V \to \mathbb{R}^K$ exists such that all distance constraints

$$\forall \{u, v\} \in E, \quad ||x_u - x_v|| = d(u, v),$$

are satisfied [4, 7], where $|| \cdot ||$ is the Euclidean norm. We focus in this extended abstract on a classical DGP application arising in the context of structural biology, where the dimension K is 3, the vertices $v \in V$ represent the atoms of a given biological molecule, and every edge $\{u, v\} \in E$ indicates whether an estimation on the distance between the two corresponding atoms is available or not [1]. Recent works have been focusing on the task of creating DGP instances for this particular application [3], in such a way to collect as much information as possible from the different available sources.

We study the specific case where the molecules that we wish to embed in the Euclidean space are *proteins*. From their primary structure, it is possible to deduce their chemical composition, i.e. we can obtain the list of atoms composing every protein, as well as the information on how each atom is connected to its neighbours by chemical bonds [5]. The chemical bonds are strong enough to allow us to obtain quite accurate values for the distance between two bonded atoms, and to estimate the distance between two atoms that are both chemically bonded to a common one.

This set of distances represents the initial source of distance information for this class of DGP instances. These distances can in practice be obtained from the parameter files of *force fields*. Force fields define a function for the potential energy of a molecule, and its parameters encapsulate information about the forces between atoms. The force fields that we are currently using in our research are the AMBER [1] and the CHARMM [2] force fields. Apart from the two types of distances mentioned above, these force fields also contain information about the so-called *var der Waals* radii, which basically imply lower bounds on the distances between pairs of atoms that are not chemically bonded.

NMR spectroscopy is an experimental technique allowing to determine information about the geometry of protein conformations from observations of the local magnetic field around atomic nuclei [10]. NMR can detect whether certain atoms in a given molecule are in close proximity, and hence it can be used to derive some approximated distances (generally represented by a lower and an upper bound) between some atom pairs. Moreover, NMR experiments provide the so-called *chemical shifts*. They describe the change in the nuclear magnetic resonance frequency of a nucleus depending on its electronic environment. By using the information

1. https://ambermd.org/AmberModels.php
2. https://www.charmm.org/archive/charmm/resources/charmm-force-fields/

about the chemical shifts, the approach TALOS+ [9], based on an Artificial Neural Network (ANN), can be employed to perform predictions on backbone dihedral (torsion) angles, which can subsequently be converted into distance information. However, we remark that this distance information is not precise (the obtained distances are represented by intervals), and that the distances cannot trivially encapsulate the information about the sign of the dihedral angle.

We point out that the collected sets of distances allow us to satisfy some special assumptions for the discretization of the DGP search space [4], and therefore to use ad-hoc software tools for the solution of the generated DGP instances [2, 6].

Current research is focusing on the possibility to include additional information in our DGP instances under the form of distance information. For example, the CHARMM force field also provides information about the some special torsion angles named *improper angles* (their name indicates the special way they are defined). While the absolute value of these angles is related to the plane angle between a triplet of consecutively bonded atoms (this information is a priori already known, see above), its sign gives us an important information about the orientation of the protein fragment. This orientation corresponds to the *chirality* [8] of the backbone part in which the improper angle is found.

This information appears to be crucial for the exploration of DGP search spaces but it is not a distance information (we have the same issue for the "proper" dihedral angles mentioned above). For this reason, future works will be aimed at devising methods for translating, when possible, these pieces of information in the form of distances. This may allow us to use the state-of-the-art DGP solvers on our DGP graphs when enriched with this additional information.

Acknowledgments. This work is partially supported by the international project MULTIBIOSTRUCT funded by the ANR French funding agency (ANR-19-CE45-0019).

Références

[1] G.M. Crippen, T.F. Havel, *Distance Geometry and Molecular Conformation*, John Wiley & Sons, 1988.

[2] D. Förster , J. Idier, L. Liberti, A. Mucherino, J-H. Lin, T.E. Malliavin, *Low-Resolution Description of the Conformational Space for Intrinsically Disordered Proteins*, Scientific Reports **12**, 19057, 16 pages, 2022.

[3] S.B. Hengeveld, T. Malliavin, J.H. Lin, L. Liberti, A. Mucherino, *A Study on the Impact of the Distance Types Involved in Protein Structure Determination by NMR*, IEEE Conference Proceedings, Computational Structural Bioinformatics Workshop (CSBW21), International Conference on Bioinformatics & Biomedicine (BIBM21), online event, 9 pages, 2021.

[4] L. Liberti, C. Lavor, N. Maculan, A. Mucherino, *Euclidean Distance Geometry and Applications*, SIAM Review **56**(1), 3–69, 2014.

[5] T.E. Malliavin, A. Mucherino, M. Nilges, *Distance Geometry in Structural Biology : New Perspectives*. In : [7], Springer, 329–350, 2013.

[6] A. Mucherino, D.S. Gonçalves, L. Liberti, J-H. Lin, C. Lavor, N. Maculan, *MD-jeep : a New Release for Discretizable Distance Geometry Problems with Interval Data*, IEEE Conference Proceedings, Federated Conference on Computer Science and Information Systems (FedCSIS20), Workshop on Computational Optimization (WCO20), Sofia, Bulgaria, 289–294, 2020.

[7] A. Mucherino, C. Lavor, L. Liberti, N. Maculan (Eds.), *Distance Geometry : Theory, Methods and Applications*, 410 pages, Springer, 2013.

[8] M. Petitjean, *Chirality in Metric Spaces*, Optimization Letters **14**, 329–338, 2020.

[9] Y. Shen, F. Delaglio, G. Cornilescu, A. Bax, *TALOS+ : a Hybrid Method for Predicting Protein Backbone Torsion Angles from NMR Chemical Shifts*, Journal of Biomolecular NMR **44**(4), 213–236, 2009.

[10] K. Wüthrich, *NMR of Proteins and Nucleic Acids*, Wiley, New York, 320 pages, 1986.