SPRINGER BRIEFS IN COMPUTER SCIENCE

Carlile Lavor
Leo Liberti
Weldon A. Lodwick
Tiago Mendonça da Costa

An Introduction to Distance Geometry applied to Molecular Geometry

Springer

SpringerBriefs in Computer Science

More information about this series at http://www.springer.com/series/10028

Carlile Lavor • Leo Liberti • Weldon A. Lodwick
Tiago Mendonça da Costa

An Introduction to Distance Geometry applied to Molecular Geometry

Springer

Carlile Lavor
University of Campinas
Campinas, São Paulo, Brazil

Leo Liberti
CNRS and École Polytechnique
Palaiseau, France

Weldon A. Lodwick
University of Colorado Denver
Denver, CO, USA

Tiago Mendonça da Costa
University of Colorado Denver
Denver, CO, USA

ISSN 2191-5768 ISSN 2191-5776 (electronic)
SpringerBriefs in Computer Science
ISBN 978-3-319-57182-9 ISBN 978-3-319-57183-6 (eBook)
DOI 10.1007/978-3-319-57183-6

Library of Congress Control Number: 2017943518

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Distance geometry (DG) is a distinct mathematical research area which includes mathematics and computer science as fundamental components. The fundamental problem of DG is to determine the spatial locations (coordinates) for a set of points, in a given geometric space, using distances between some of them.

DG is considered to have originated in 1928, when Menger [62] characterized several geometric concepts using the idea of distance [16]. However, only with the results of Blumenthal [10], in 1953, did the topic become a new area of knowledge known as DG.

The main challenge of DG at that time was to find necessary and sufficient conditions in order to decide if a given matrix is a distance matrix D. That is, decide whether or not a given matrix D is a symmetric matrix such that there is an integer number K and a set of points in $\mathbb{R}^K$, where the Euclidean distances between these points are equal to the entries of the matrix D [51]. Note that, in this case, all distances are considered known.

To the best of our knowledge, the first explicit mention of the fundamental problem of DG delineated above, where not all distances are known, was given by Yemini [81], in 1978. In this case, the problem may be harder.

Another important moment in the history of DG is related to its application to the calculation of molecular structures, with the 1988 publication of Crippen and Havel's book [15], considered pioneers of DG in the analysis of protein structures. More information on the DG history can be found in [52].

The first edited book fully dedicated to DG was published by Springer in 2013 [67]. The book brought together different applications and researchers in DG. In the same year, in June 2013, the first international workshop dedicated to DG was held with speakers from various international institutions (Princeton University, IBM TJ Watson Research Center, University of Cambridge, École Polytechnique, Institut Pasteur, École Normale Supérieure, SUTD-MIT International Design Center). The event also had the support of several international scientific societies and universities, indicating the importance of DG in many areas of knowledge (more details at http://dga2013.icomp.ufam.edu.br/).

The interest in DG as a topic of research arises from the wealth and diversity of its applications, in addition to its mathematical depth and beauty. Recent surveys on DG highlighting the theory and applications can be found in [9, 18, 57]. For example, applications can be found in problems from astronomy, biochemistry, statistics, nanotechnology, robotics, and telecommunications.

In astronomy, the problem is related to the determination of star positions using information about the distances between some of them [60]. In biochemistry, the problem appears in the determination of three-dimensional structures of protein molecules using the information obtained from nuclear magnetic resonance (NMR) experiments. In statistics, there are problems related to visualization of data [22] and dimensionality reduction [50]. In these cases, points and distances are given in a high dimensional space $\mathbb{R}^n$ and the problem is how represent them in a lower dimension, say $\mathbb{R}^2$ or $\mathbb{R}^3$, in order to have a visual idea of the data. This application is also linked to a current topic of research called *Big Data* [2, 61]. In nanotechnology, the problem is similar to the problem in biochemistry, but on a "nano" scale [21, 39]. There is a direct relationship between the application to robotics and the calculations related to molecular geometry [23, 70]. That is, given a set of robotic arm lengths (distances), the problem is to find the locus of points that the robot arm can reach [68]. In telecommunications, the problem is related to the positioning of a wireless sensor network, where the distances can be estimated by the amount of power necessary for performing peer-to-peer sensor communication. The further the sensors, the more power is necessary. Since both sensors know how much power they used, both sensors can compute their distance. An example of this is for router positioning [24, 81].

The theoretical nature and the wide variety of applications have resulted in DG becoming its own research area in applied mathematics, which includes fundamental concepts from mathematics (measures, norms, geometry, optimization, combinatorics, graph theory, symmetry, uncertainty) and computer science (algorithms, solvability, complexity).

Campinas, São Paulo, Brazil	Carlile Lavor
Palaiseau, France	Leo Liberti
Denver, CO, USA	Weldon A. Lodwick
Denver, CO, USA	Tiago Mendonça da Costa
March 2017	

Acknowledgements

We would like to thank the Brazilian research agencies CNPq, CAPES, and FAPESP for their financial support, and Prof. Paulo J. S. Silva (from University of Campinas) and Prof. Douglas S. Gonçalves (from Federal University of Santa Catarina) for their important comments and suggestions that improve the final version of the book.

We also thank the Brazilian Society of Computational and Applied Mathematics (SBMAC) for the opportunity to give a Distance Geometry course that resulted in the first version of this monograph in Portuguese (http://www.sbmac.org.br/p_notas.php).

Contents

Chapter 1
Introduction

This monograph introduces distance geometry, based on problems related to 3D protein structure calculations using distance information provided by Nuclear Magnetic Resonance (NMR) experiments. Our presentation is as short as possible, with exercises to help the reader to better understand the contents.

The text is based on the combinatorial structure of distance geometry problems, differently from the classical approach, which focuses on continuous methods. This discrete approach is important for the understanding of the main concepts involved and provides a new way to consider the problem. Curiously, this combinatorial view arose in the quantum computing context, when Grover's algorithm [3, 33, 42] was proposed as a method to solve a distance geometry problem related to the calculation of molecular structures [41].

1.1 Notation and Basic Concepts

It is assumed that an elementary knowledge of analytic geometry will be sufficient to follow the text. Obviously, the reader will also need to have some familiarity with the mathematical language involving basic concepts from logic, sets, and functions. Next, we list the main concepts that will be used in this text with the associated notation.

- **Sets**
  - $x \in A$ means that x is an element of set A.
  - $A \subset B$ means that the set A is contained in B, that is, all elements in A are elements of B.
  - $A \cap B = \{x : x \in A \text{ and } x \in B\}$ is the set formed by the elements that are both in A and B.

C. Lavor et al., *An Introduction to Distance Geometry applied to Molecular Geometry*, SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-57183-6_1

  - $A \cup B = \{x : x \in A \text{ or } x \in B\}$ is the set formed by the elements that are in A or B, including the elements which are in $A \cap B$.
  - $A - B = \{x : x \in A \text{ and } x \notin B\}$ is the set A, excluding the elements in B which are in A.
  - The set A is countably infinite if and only if there exists a bijection $f : \mathbb{N} \to A$, that is, a one-to-one correspondence between the elements in A and the positive integer numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$.
  - The set A is finite if and only if there exists a bijection $f : \{1, 2, \ldots, n\} \to A$, where $n \in \mathbb{N}$. The cardinality of A, which is denoted by $|A|$, is the number of elements in A.
  - The set A is uncountable if it is infinite but it is not countably infinite.

- **Vectors and Matrices**
  - A vector $x \in \mathbb{R}^n$ will be denoted by a column matrix. For example, $x \in \mathbb{R}^2$ will be written as $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, where $x_1, x_2 \in \mathbb{R}$.
  - $\mathbb{R}^{m \times n}$ is the set of matrices with m lines and n columns whose entries are real numbers.
  - Given $M \in \mathbb{R}^{m \times n}$, the transpose matrix $M^T \in \mathbb{R}^{n \times m}$ is the real matrix obtained from M by exchanging rows by columns.
  - Given $M \in \mathbb{R}^{m \times m}$, its inverse matrix, denoted by $M^{-1} \in \mathbb{R}^{m \times m}$, exists if and only if $MM^{-1} = M^{-1}M = I$, where I is the identity matrix. For example, if $I \in \mathbb{R}^{2 \times 2}$, then $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. If M^{-1} exists, we say that M is invertible. Recall that the product between matrices is not commutative, that is, there are matrices A, B, such that $AB \neq BA$.
  - Given $M \in \mathbb{R}^{m \times n}$, with $m \leq n$, we say that M has full rank if and only if there exists a submatrix $M' \in \mathbb{R}^{m \times m}$, where M' is invertible.
  - Given $x, y \in \mathbb{R}^n$, the inner product between x and y, denoted by $x \cdot y$, is defined by $x \cdot y = x_1 y_1 + \cdots + x_n y_n$.
  - Given $x \in \mathbb{R}^n$, the Euclidean norm of x, denoted by $||x||$, is defined by $||x|| = \sqrt{x \cdot x} = \sqrt{x_1^2 + \cdots + x_n^2}$.

- **Graphs**
  - Graph theory is a discipline which intersects mathematics and computer science [34]. Here we present some basic definitions that we need for our development.
  - Given a finite set V and another set E, which is formed by unordered pairs of elements in V, we have a graph $G = (V, E)$, where V is the set of vertices and E is the set of edges. We can represent a graph in the plane $\mathbb{R}^2$ using points as the vertices and line segments (or arcs) as the edges, as illustrated below.
  - A graph $G = (V, E)$ is simple if and only if it has no multiple edges or $\{a, b\} \in E \Rightarrow a \neq b$ (see Figs. 1.1 and 1.2).

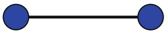


Fig. 1.1 Simple graph

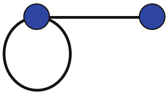


Fig. 1.2 Non-simple graph

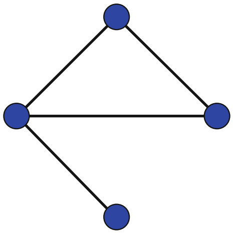


Fig. 1.3 Connected graph

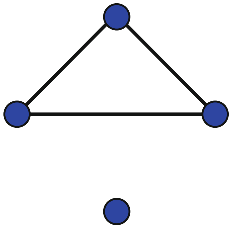


Fig. 1.4 Disconnected graph

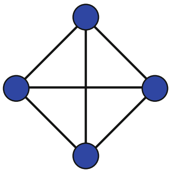


Fig. 1.5 Complete graph

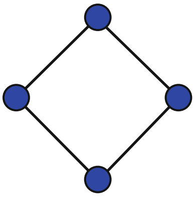


Fig. 1.6 A graph that is not complete

- A graph $G = (V, E)$ is connected if and only if it is not possible to separate the set of vertices in two non-empty sets A and B, $V = A \cup B$, such that there is no edges $\{a, b\} \in E$ where $a \in A$ and $b \in B$ (see Figs. 1.3 and 1.4).
- A graph $G = (V, E)$ is complete if and only if the set E contains all possible pairs, that is, for all $a, b \in V, a \neq b \Rightarrow \{a, b\} \in E$ (see Figs. 1.5 and 1.6).
- A clique in a graph $G = (V, E)$ is another graph $G' = (V', E')$, where $V' \subset V$, $E' \subset E$, and $G' = (V', E')$ is complete (see Figs. 1.7, 1.8, and 1.9).
- When we associate each edge of a graph with a real number, that is, a function $d : E \to \mathbb{R}$ is given, we have a graph with weights on the edges (see Fig. 1.10).

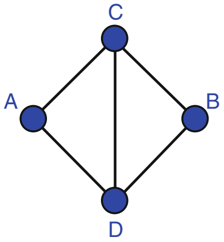


Fig. 1.7 Graph *G*

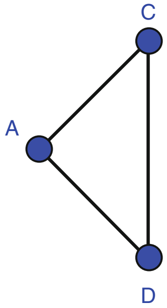


Fig. 1.8 3-Clique in *G*

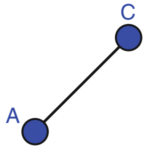


Fig. 1.9 2-Clique in *G*

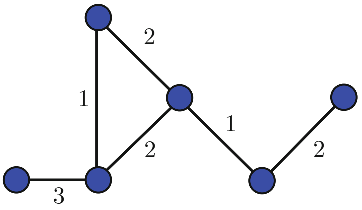


Fig. 1.10 Edge-weighted graph

1.2 Outline

The remainder of the book is the following. In Chap. 2, we define the basic problem of Distance Geometry: the DGP (Distance Geometry Problem). In Chap. 3, we present the continuous approach to the DGP and introduce some ideas related to the combinatorial approach. Chapters 4 and 5 consider two discrete versions of the DGP and Chap. 6 explains how Distance Geometry can be used to model problems in Molecular Geometry associated to 3D protein structure calculations using NMR data. Chapter 7 ends with some conclusions.

Chapter 2
The Distance Geometry Problem (DGP)

2.1 Definition of the DGP

The fundamental problem of DG, as we have previously stated, is to determine all the coordinates of a set points, in a given geometric space, for which some of the distances are known. Depending on the application, these points can represent stars, reachable points for a robot arm, atoms, or people. Each one of these objects can be represented by a vertex of a graph, and if the distance between them is known, we have an edge connecting the correspondent vertices. Formally, we have the following definition of the Distance Geometry Problem (DGP) [57].

Definition 2.1 (DGP) Given a integer $K > 0$ and a simple connected graph $G = (V, E)$ with weights on the edges given by $d : E \to (0, \infty)$, find a function $x : V \to \mathbb{R}^K$ such that

$$\forall \{u, v\} \in E, \ \|x(u) - x(v)\| = d(u, v). \tag{2.1}$$

Remark 2.1 The norm of (2.1) is general and will depend on the application. This monograph uses the Euclidean norm.

A solution to the DGP associates each vertex of G to a point in $\mathbb{R}^K$ satisfying Eq. (2.1). That is, we wish to position the vertices $u, v \in V$ such that, for $\{u, v\} \in E$, we have situated them in $\mathbb{R}^K$ so that the calculated distance $\|x(u)-x(v)\|$ is the given value $d(u, v)$. The function x is called a *realization* of G. A realization of a graph is a "representation" of its vertices in some Euclidean space $\mathbb{R}^K$. Note that the dimension K and the graph G are inputs/data of the problem. Some DGP variants have the dimension K as part of the problem [57]. This monograph, however, assumes that the dimension K is given a priori.

A realization that satisfies all Eq. (2.1) is a *valid realization*. In order to simplify the notation, we will use x_u, x_v instead of $x(u)$, $x(v)$, and d_{uv} instead of $d(u, v)$.

C. Lavor et al., *An Introduction to Distance Geometry applied to Molecular Geometry*, SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-57183-6_2

The focus here is on the cases $K = 2$ and $K = 3$. However, all results can be extended to $\mathbb{R}^K$ [57].

Exercise 2.1 Can there exist more than one solution of a DGP? Can the solution set be empty?

Exercise 2.2 When we "draw" a graph on paper, are we solving a DGP?

We obtain the following system of equations when we use the Euclidean norm in the definition of the DGP with $K = 2$:

$$\sqrt{(x_{u1} - x_{v1})^2 + (x_{u2} - x_{v2})^2} = d_{uv} \ \forall \{u, v\} \in E, \tag{2.2}$$

where $x_u^T = (x_{u1}, x_{u2})$ and $x_v^T = (x_{v1}, x_{v2})$. Thus, we have a system with $2|V|$ variables and $|E|$ equations. From (2.2), by squaring both sides, we immediately derive:

$$(x_{u1} - x_{v1})^2 + (x_{u2} - x_{v2})^2 = d_{uv}^2 \ \forall \{u, v\} \in E. \tag{2.3}$$

Trying to solve the system (2.1) or its associated quadratic system (2.3) as a closed formula is, in general, impossible [7]. Solving the problem numerically also presents difficulties [57].

Exercise 2.3 Consider a DGP with $K = 2, V = \{u, v, r\}, E = \{\{u, v\}, \{v, r\}\}$, and $d_{uv} = d_{vr} = 1$. Solve the problem graphically.

Exercise 2.4 Considering the previous exercise, what would be the solution if we add $\{u, r\}$ to E with $d_{ur} = 1$?

Before we consider solution methods to solve the DGP, we will discuss two important aspects of the problem: (i) the cardinality of the solution set and (ii) the complexity of the problem.

2.2 Number of Solutions of the DGP

What is the importance of knowing the number of solutions of a DGP? Is it possible to have this information before we solve the problem? Besides its theoretical importance, the cardinality of the DGP solution set may help us to solve the problem, since we know how many solutions are being sought. These questions will be discussed more fully in Chap. 5.

We saw from the first two exercises that the number of solutions of a DGP can be infinite. However, does this always happen? We know that, given three points x_u, x_v, x_w in $\mathbb{R}^2$, the triangle inequality

$$d_{uw} \leq d_{uv} + d_{vw}$$

must be satisfied, where d_{uv}, d_{vw}, d_{uw} are the distances between the given points. This means that if we add the edge $\{u, r\}$ to E (in the last exercise), the problem does not have a solution if the triangle inequality is not satisfied.

It is also clear that once we have a solution, there will be an unaccountably infinite number of other solutions by simply rotating and/or translating the given solution. However, excluding rotations, translations, and reflections, the exercises suggest that the set of solutions of a DGP can be of three kinds:

- empty,
- finite,
- unaccountably infinite.

There is one more case to consider. Is it possible to have a DGP with a countably infinite number of solutions? It turns out that this case is impossible, but the proof is not simple enough to be presented here [8].

Exercise 2.5 Consider a DGP with $K = 2, V = \{u, v, r, s\}$, $E = \{\{u, v\}, \{u, r\}, \{v, r\}, \{v, s\}\}$ and $d_{uv} = d_{ur} = d_{vr} = d_{vs} = 1$. Excluding rotations, translations, and reflections, how many solutions exist?

Exercise 2.6 Considering the previous exercise, how many solutions will exist if we add $\{u, s\}$ to E, with $d_{us} = \sqrt{2}$?

The two exercises above illustrate the fact that the addition of a single edge can engender a change from "uncountably many" to "finitely many" solutions. This is an evidence (not a proof) that the DGP cannot have countably infinitely many solutions.

2.3 Complexity of the DGP

The focus of this section is to give an intuition of the computational difficulty we face when solving a DGP (formally, this is investigated in computational complexity theory). Let us first consider the DGP whose associated graph is complete. To this end, consider a DGP with $K = 1$, $V = \{u, v, r\}$, $E = \{\{u, v\}, \{u, r\}, \{v, r\}\}$, $d_{uv} = d_{vr} = 1$, and $d_{ur} = 2$. If we fix $x_u = 0$ and $x_v = 1$, we have

$$\begin{aligned} \|x_r - x_u\| &= 2 \\ \|x_r - x_v\| &= 1. \end{aligned}$$

Squaring both terms of equalities, we have

$$\begin{aligned} x_r^2 - 2x_r x_u + x_u^2 &= 4 \\ x_r^2 - 2x_r x_v + x_v^2 &= 1. \end{aligned}$$

Subtracting one equation from the other, we obtain

$$-2x_rx_u + 2x_rx_v + x_u^2 - x_v^2 = 3 \Rightarrow 2x_r(x_v - x_u) = x_v^2 - x_u^2 + 3.$$

Using the fixed values for x_u and x_v, we get

$$2x_r = 4 \Rightarrow x_r = 2.$$

We could have solved this problem by simply drawing the graph. However, the interesting part of this procedure is that it can be generalized for $\mathbb{R}^K$. Let us see what happens in $\mathbb{R}^2$.

Consider a DGP with $K = 2$,

$$V = \{u, v, r, s\} \text{ and } E = \{\{u, v\}, \{u, r\}, \{u, s\}, \{v, r\}, \{v, s\}, \{r, s\}\}.$$

Assume that u, v, r are fixed, that is, we can find $x_u, x_v, x_r \in \mathbb{R}^2$ such that $\|x_u - x_v\| = d_{uv}$, $\|x_u - x_r\| = d_{ur}$, $\|x_v - x_r\| = d_{vr}$. Given these three points in $\mathbb{R}^2$, we can construct a quadratic system to obtain the coordinates of x_s in the following way:

$$\begin{aligned}\|x_s - x_u\| &= d_{us}\\ \|x_s - x_v\| &= d_{vs}\\ \|x_s - x_r\| &= d_{rs}.\end{aligned}$$

Squaring both terms of equalities,

$$\begin{aligned}\|x_s\|^2 - 2x_s \cdot x_u + \|x_u\|^2 &= d_{us}^2\\ \|x_s\|^2 - 2x_s \cdot x_v + \|x_v\|^2 &= d_{vs}^2\\ \|x_s\|^2 - 2x_s \cdot x_r + \|x_r\|^2 &= d_{rs}^2,\end{aligned}$$

and subtracting the first equation from the other two, we obtain

$$\begin{aligned}2(x_v - x_u) \cdot x_s &= \|x_v\|^2 - \|x_u\|^2 + d_{us}^2 - d_{vs}^2\\ 2(x_r - x_u) \cdot x_s &= \|x_r\|^2 - \|x_u\|^2 + d_{us}^2 - d_{rs}^2.\end{aligned}$$

Thus, we have a linear system

$$Ax = b,$$

where

$$A = 2\begin{bmatrix} x_{v1} - x_{u1} & x_{v2} - x_{u2}\\ x_{r1} - x_{u1} & x_{r2} - x_{u2}\end{bmatrix},$$

$$b = \begin{bmatrix} \|x_v\|^2 - \|x_u\|^2 + d_{us}^2 - d_{vs}^2 \\ \|x_r\|^2 - \|x_u\|^2 + d_{us}^2 - d_{rs}^2 \end{bmatrix},$$

and

$$x = \begin{bmatrix} x_{s1} \\ x_{s2} \end{bmatrix}.$$

If the matrix A is invertible, then we have only one solution x^* given by

$$x^* = A^{-1}b. \tag{2.4}$$

Remark 2.2 Note that supposing that the matrix A is invertible, the solution x^* is obtained for any values of d_{us}, d_{vs}, d_{rs}. What is the connection of this fact with the solution of the original quadratic system?

If A has no inverse, what do we do? We can, for example, fix three different vertices of $\{u, v, r, s\}$ in order to obtain an invertible matrix and hence the position of the fourth vertex. Therefore, if the graph G of a DGP is complete and we assume the existence of a solution, we can obtain a realization of G by solving a sequence of linear systems, considering that all the associated matrices are invertible.

Exercise 2.7 In Exercise 2.5, suggest a method for fixing the positions of u, v, r. Does it generalize to $\mathbb{R}^3$?

Exercise 2.8 In Exercise 2.5, what conditions can we impose on x_u, x_v, x_r to guarantee that the associated linear system has a solution?

Exercise 2.9 In Exercise 2.5, if the associated linear system has a solution, can we guarantee that we have a solution to the original quadratic system?

Exercise 2.10 Still considering Exercise 2.5, if none of possible choices for the three first vertices produce an invertible matrix, does this mean that the DGP has no solution?

If we require that all the matrices associated to a DGP with complete graph possess inverses, the problem has a unique solution (modulo rotations, translations, and reflections) that can be obtained at a computational cost proportional to $|V|$ [20]. However, in the majority of applications, we do not have complete graphs. So, the strategy discussed above may yield just a partial realization of the graph.

Before continuing our exploration of the DGP, we state a theoretical result pertaining to the computational complexity of the DGP [72].

Theorem 2.1 *The DGP is NP-Hard for all $K \in \mathbb{N}$, and in particular it is NP-Complete for $K = 1$.*

This means that any algorithm capable of solving the DGP is very likely to run (in the worst case) in a number of steps which is exponential in the size of the memory used to store the instance data, that is, G and d.

Exercise 2.11 To solve an instance of the DGP in linear time, does the associated graph necessarily need to be complete?

Exercise 2.12 Can we have just a unique solution (modulo rotations, translations and reflections) to the DGP even though the graph is not complete?

2.4 DGP Instances

It is important to be able to generate instances/examples of DGPs in order to test algorithms and to analyze the relationship between the graph of a DGP and its set of solutions. This section presents a method for the reader to generate example problems of a DGP for $K = 3$.

The procedure given here stems from the calculation of 3D molecular structures. To simplify the process without making the generated instance easy to solve, we consider a sequence of covalently connected atoms, denoted by $1, \ldots, n$. That is, each atom is connected to only two others, except the first and the last of the sequence.

We will use a Cartesian coordinate system $x_1, \ldots, x_n \in \mathbb{R}^3$ to define a spatial structure of our molecule in terms of an *internal coordinate system* [77], given by the *lengths of the covalent bonds* $d_{1,2}, \ldots, d_{n-1,n}$, by the *planar angles* $\theta_{1,3}, \ldots, \theta_{n-2,n}$ (formed by three consecutive atoms), and by the *torsion angles* $\omega_{1,4}, \ldots, \omega_{n-3,n}$ (formed by four consecutive atoms). Each torsion angle $\omega_{i-3,i}$ is, in fact, the angle between the normals of the planes defined by the atoms $i-3, i-2, i-1$ and $i-2, i-1, i$, respectively (see Fig. 2.1).

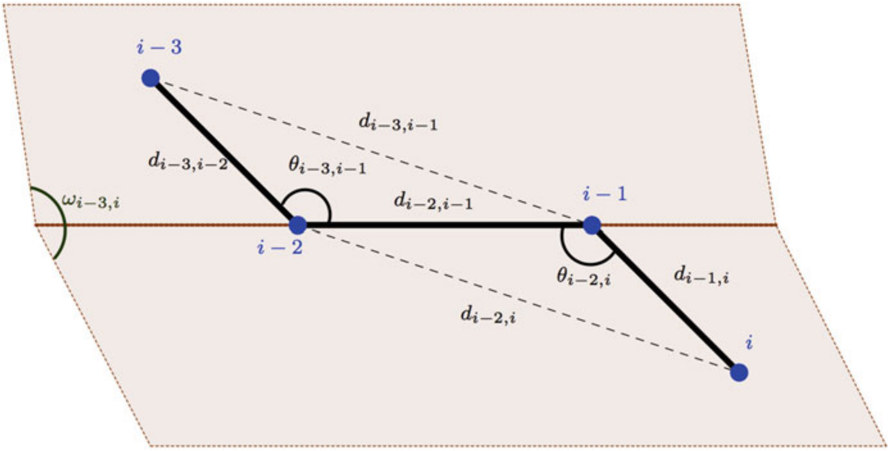


Fig. 2.1 Internal coordinates

We now fix the lengths of the covalent bonds (for example, $d_{i-1,i} = 1.526$) and the values of the planar angles (for example, $\theta_{i-2,i} = 1.91$ radians). In this way, except for rotations, translations, and reflections, a structure for our molecule will be determined by the torsion angles $\omega_{1,4}, \ldots, \omega_{n-3,n}$, each of which can vary in the interval $[0, 2\pi]$.

A way to proceed is to choose randomly values for $\omega_{i-3,i} \in [0, 2\pi]$, as well as pairs of points i, j whose Euclidean distances d_{ij} are smaller than a given value. To simulate DGP instances associated with the calculation of molecular structures using distance information provided by NMR experiments, we can choose pairs of points i, j for which $d_{ij} \leq 5$ [40] (see Chap. 6).

However, how do we determine the pairs of points i, j without knowing the distances d_{ij}? For this, we need to obtain the Cartesian coordinates from the internal coordinates, as follows:

$$\begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ 1 \end{bmatrix} = B_1 B_2 \cdots B_i \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \ \forall i = 1, \ldots, n,$$

where

$$B_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \ B_2 = \begin{bmatrix} -1 & 0 & 0 & -d_{1,2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$B_3 = \begin{bmatrix} -\cos\theta_{1,3} & -\sin\theta_{1,3} & 0 & -d_{2,3}\cos\theta_{1,3} \\ \sin\theta_{1,3} & -\cos\theta_{1,3} & 0 & d_{2,3}\sin\theta_{1,3} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

and

$$B_i = \begin{bmatrix} -\cos\theta_{i-2,i} & -\sin\theta_{i-2,i} & 0 & -d_{i-1,i}\cos\theta_{i-2,i} \\ \sin\theta_{i-2,i}\cos\omega_{i-3,i} & -\cos\theta_{i-2,i}\cos\omega_{i-3,i} & -\sin\omega_{i-3,i} & d_{i-1,i}\sin\theta_{i-2,i}\cos\omega_{i-3,i} \\ \sin\theta_{i-2,i}\sin\omega_{i-3,i} & -\cos\theta_{i-2,i}\sin\omega_{i-3,i} & \cos\omega_{i-3,i} & d_{i-1,i}\sin\theta_{i-2,i}\sin\omega_{i-3,i} \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

for $i = 4, \ldots, n$.

Using above matrices and fixing the values $d_{1,2}, d_{2,3}, \theta_{1,3}$, the coordinates of the first three atoms are given by:

$$x_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

$$x_2 = \begin{bmatrix} -d_{1,2} \\ 0 \\ 0 \end{bmatrix},$$

$$x_3 = \begin{bmatrix} -d_{1,2} + d_{2,3} \cos \theta_{1,3} \\ d_{2,3} \sin \theta_{1,3} \\ 0 \end{bmatrix}.$$

Note that we are using matrices in $\mathbb{R}^{4\times 4}$ to generate points in $\mathbb{R}^3$. These are related to *homogeneous coordinates*, which are very useful in computer graphics [29].

Exercise 2.13 Since 3D molecular structures can be described by Cartesian coordinates or internal coordinates, what is the difference in using one or the other coordinate system to generate a DGP instance?

Exercise 2.14 Is it possible to use the matrices above to generate DGP instances for $K = 2$?

Chapter 3
From Continuous to Discrete

3.1 Continuous Optimization and the DGP

One approach that has been used to solve the DGP is to represent it as a continuous optimization problem [59]. To understand it, we consider a DGP with $K = 2$, $V = \{u, v, s\}$, $E = \{\{u, v\}, \{v, s\}\}$, where the associated quadratic system is

$$\begin{aligned}(x_{u1} - x_{v1})^2 + (x_{u2} - x_{v2})^2 &= d_{uv}^2\\(x_{v1} - x_{s1})^2 + (x_{v2} - x_{s2})^2 &= d_{vs}^2,\end{aligned}$$

which can be rewritten as

$$\begin{aligned}(x_{u1} - x_{v1})^2 + (x_{u2} - x_{v2})^2 - d_{uv}^2 &= 0\\(x_{v1} - x_{s1})^2 + (x_{v2} - x_{s2})^2 - d_{vs}^2 &= 0.\end{aligned}$$

Consider the function $f : \mathbb{R}^6 \to \mathbb{R}$, defined by

$$\begin{aligned}f(x_{u1}, x_{u2}, x_{v1}, x_{v2}, x_{s1}, x_{s2}) = &\left((x_{u1} - x_{v1})^2 + (x_{u2} - x_{v2})^2 - d_{uv}^2\right)^2\\&+ \left((x_{v1} - x_{s1})^2 + (x_{v2} - x_{s2})^2 - d_{vs}^2\right)^2.\end{aligned}$$

It is not hard to realize that the solution $x^* \in \mathbb{R}^6$ of the associated DGP can be found by solving the following problem:

$$\min_{x \in \mathbb{R}^6} f(x). \tag{3.1}$$

That is, we wish to find the point $x^* \in \mathbb{R}^6$ which attains the smallest value of f.

C. Lavor et al., *An Introduction to Distance Geometry applied to Molecular Geometry*, SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-57183-6_3

Exercise 3.1 In the problem (3.1), is it possible to say what is the smallest valued of f? Is this result valid for any DGP?

Exercise 3.2 Still considering the same problem above (3.1), what is the difference between solving the quadratic system

$$(x_{u1} - x_{v1})^2 + (x_{u2} - x_{v2})^2 - d_{uv}^2 = 0$$
$$(x_{u1} - x_{s1})^2 + (x_{v2} - x_{s2})^2 - d_{vs}^2 = 0$$

and solving the equation

$$f(x) = 0, x \in \mathbb{R}^6,$$

where f is given by

$$f(x_{u1}, x_{u2}, x_{v1}, x_{v2}, x_{s1}, x_{s2}) = \left((x_{u1} - x_{v1})^2 + (x_{u2} - x_{v2})^2 - d_{uv}^2\right)^2 + \left((x_{u1} - x_{s1})^2 + (x_{v2} - x_{s2})^2 - d_{vs}^2\right)^2?$$

Thus, we can think of the DGP as a minimization problem. However, the optimization approach for the DGP has a difficulty in that the function to be minimized (3.1) has many local minima and we wish to find a global minimum [26] (see Fig. 3.1).

In fact, the number of local minima may increase exponentially with the size of the problem, which is determined by the number of vertices of the associated graph [57], further complicating the minimization problem.

Exercise 3.3 Is it possible that there exists more than one global minimum for the DGP optimization problem?

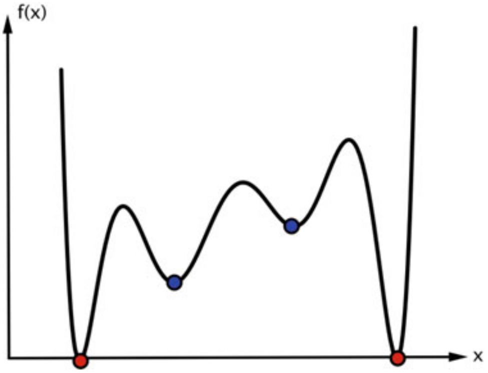


Fig. 3.1 Local and global minima

Exercise 3.4 If there is more than one global minimum, does the number of global minima may increase exponentially with the size of the problem in the same way that local minima do?

3.2 Finiteness of the DGP

Suppose that the solution set of a DGP is non-empty. We already know that the number of solutions is either uncountable or finite (modulo rotations, translations, and reflections). In the finite case, besides applying the classical optimization methods, we can exploit the structure of the associated graph which defines the problem and perhaps come up with a different approach [46, 54].

However, before studying special problem structures, how can we tell if the solution set is either finite or uncountable? This section analyzes the conditions that ensure the finiteness of the solutions for a DGP.

Let us consider the same problem of Sect. 2.3 with $K = 2$, but here we change the dimension K to 3. So, we have a DGP with $K = 3$, $V = \{u, v, r, s\}$, $E = \{\{u, v\}, \{u, r\}, \{u, s\}, \{v, r\}, \{v, s\}, \{r, s\}\}$. Fixing the coordinates of the first three vertices u, v, r, which we can do by using the matrices of Sect. 2.4, we obtain the same quadratic system:

$$\begin{aligned}
\|x_s - x_u\|^2 &= d_{us}^2,\\
\|x_s - x_v\|^2 &= d_{vs}^2,\\
\|x_s - x_r\|^2 &= d_{rs}^2.
\end{aligned}$$

Performing the calculations and subtracting the first equation from the others as before, we obtain

$$\begin{aligned}
2(x_v - x_u) \cdot x_s &= \|x_v\|^2 - \|x_u\|^2 + d_{us}^2 - d_{vs}^2,\\
2(x_r - x_u) \cdot x_s &= \|x_r\|^2 - \|x_u\|^2 + d_{us}^2 - d_{rs}^2.
\end{aligned}$$

So far, no difference can be noticed. However, if we obtain the explicit associated linear system, we have:

$$\begin{bmatrix} x_{v1} - x_{u1} & x_{v2} - x_{u2} & x_{v3} - x_{u3} \\ x_{r1} - x_{u1} & x_{r2} - x_{u2} & x_{r3} - x_{u3} \end{bmatrix} \begin{bmatrix} x_{s1} \\ x_{s2} \\ x_{s3} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \|x_v\|^2 - \|x_u\|^2 + d_{us}^2 - d_{vs}^2 \\ \|x_r\|^2 - \|x_u\|^2 + d_{us}^2 - d_{rs}^2 \end{bmatrix}.$$

We no longer have a 2 × 2 matrix, but a 2 × 3 matrix, since we have $K = 3$.

The above system can be written as

$$\begin{bmatrix} x_{v1}-x_{u1} & x_{v2}-x_{u2} \\ x_{r1}-x_{u1} & x_{r2}-x_{u2} \end{bmatrix} \begin{bmatrix} x_{s1} \\ x_{s2} \end{bmatrix} + \begin{bmatrix} x_{v3}-x_{u3} \\ x_{r3}-x_{u3} \end{bmatrix} \begin{bmatrix} x_{s3} \end{bmatrix}$$
$$= \frac{1}{2} \begin{bmatrix} \|x_v\|^2 - \|x_u\|^2 + d_{us}^2 - d_{vs}^2 \\ \|x_r\|^2 - \|x_u\|^2 + d_{us}^2 - d_{rs}^2 \end{bmatrix}.$$

If we suppose that the matrix $\begin{bmatrix} x_{v1}-x_{u1} & x_{v2}-x_{u2} \\ x_{r1}-x_{u1} & x_{r2}-x_{u2} \end{bmatrix}$ is invertible, we obtain that

$$\begin{bmatrix} x_{s1} \\ x_{s2} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} x_{v1}-x_{u1} & x_{v2}-x_{u2} \\ x_{r1}-x_{u1} & x_{r2}-x_{u2} \end{bmatrix}^{-1} \begin{bmatrix} \|x_v\|^2 - \|x_u\|^2 + d_{us}^2 - d_{vs}^2 \\ \|x_r\|^2 - \|x_u\|^2 + d_{us}^2 - d_{rs}^2 \end{bmatrix}$$
$$- \begin{bmatrix} x_{v1}-x_{u1} & x_{v2}-x_{u2} \\ x_{r1}-x_{u1} & x_{r2}-x_{u2} \end{bmatrix}^{-1} \begin{bmatrix} x_{v3}-x_{u3} \\ x_{r3}-x_{u3} \end{bmatrix} \begin{bmatrix} x_{s3} \end{bmatrix}.$$

This implies that we no longer have only one solution for the linear system! That is, for each value for $x_{s3} \in \mathbb{R}$, we obtain values for x_{s1} and x_{s2}. Thus, in order to obtain a solution of our DGP, we must return to the associated quadratic system, choose one of the equations (for example, $\|x_s - x_u\|^2 = d_{us}^2$), and solve it by using the solution of the linear system above.

Geometrically, we have the intersection between a line, given by the parametric equation in x_{s3},

$$\begin{bmatrix} x_{s1} \\ x_{s2} \end{bmatrix} = A - B \begin{bmatrix} x_{s3} \end{bmatrix},$$

where

$$A = \frac{1}{2} \begin{bmatrix} x_{v1}-x_{u1} & x_{v2}-x_{u2} \\ x_{r1}-x_{u1} & x_{r2}-x_{u2} \end{bmatrix}^{-1} \begin{bmatrix} \|x_v\|^2 - \|x_u\|^2 + d_{us}^2 - d_{vs}^2 \\ \|x_r\|^2 - \|x_u\|^2 + d_{us}^2 - d_{rs}^2 \end{bmatrix},$$
$$B = \begin{bmatrix} x_{v1}-x_{u1} & x_{v2}-x_{u2} \\ x_{r1}-x_{u1} & x_{r2}-x_{u2} \end{bmatrix}^{-1} \begin{bmatrix} x_{v3}-x_{u3} \\ x_{r3}-x_{u3} \end{bmatrix},$$

and a sphere, given by

$$\|x_s - x_u\|^2 = d_{us}^2, \tag{3.2}$$

resulting in three possibilities (Fig. 3.2):

- Empty set (the line does not intersect the sphere),
- Only one point (the line is tangent to sphere),
- Two points (the line is a secant of the sphere).

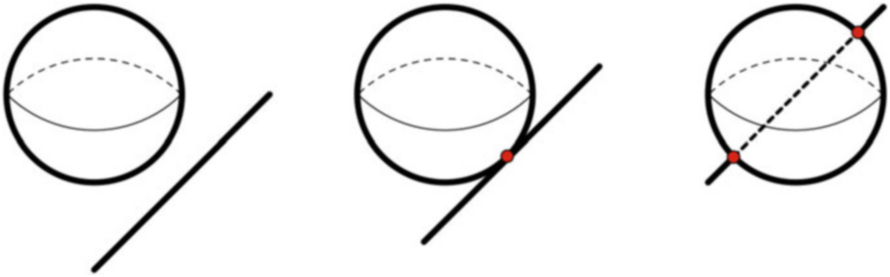


Fig. 3.2 Intersection of a line and a sphere

Exercise 3.5 Show the equivalence between the original quadratic system,

$$\|x_s - x_u\|^2 = d_{us}^2,$$
$$\|x_s - x_v\|^2 = d_{vs}^2,$$
$$\|x_s - x_r\|^2 = d_{rs}^2,$$

and the new system, given by

$$\left\| \begin{bmatrix} x_{s1} & x_{s2} & x_{s3} \end{bmatrix}^T - \begin{bmatrix} x_{u1} & x_{u2} & x_{u3} \end{bmatrix}^T \right\|^2 = d_{us}^2,$$

$$\begin{bmatrix} x_{s1} \\ x_{s2} \end{bmatrix} = \begin{bmatrix} x_{v1} - x_{u1} & x_{v2} - x_{u2} \\ x_{r1} - x_{u1} & x_{r2} - x_{u2} \end{bmatrix}^{-1}$$
$$\times \left\{ \frac{1}{2} \begin{bmatrix} \|x_v\|^2 - \|x_u\|^2 + d_{us}^2 - d_{vs}^2 \\ \|x_r\|^2 - \|x_u\|^2 + d_{us}^2 - d_{rs}^2 \end{bmatrix} - \begin{bmatrix} x_{v3} - x_{u3} \\ x_{r3} - x_{u3} \end{bmatrix} \begin{bmatrix} x_{s3} \end{bmatrix} \right\},$$

where the variables are $x_{s1}, x_{s2}, x_{s3} \in \mathbb{R}$.

Exercise 3.6 If the matrix $\begin{bmatrix} x_{v1} - x_{u1} & x_{v2} - x_{u2} \\ x_{r1} - x_{u1} & x_{r2} - x_{u2} \end{bmatrix}$ is not invertible, or if we can not select an invertible matrix 2 × 2 from the original matrix

$$\begin{bmatrix} x_{v1} - x_{u1} & x_{v2} - x_{u2} & x_{v3} - x_{u3} \\ x_{r1} - x_{u1} & x_{r2} - x_{u2} & x_{r3} - x_{u3} \end{bmatrix},$$

does this mean that we can not ensure the finiteness of the solution set?

The discussion above suggests that there are at least two important aspects to the issue of finiteness of the DGP solution set:

a For each vertex $s \in V$ that we need to process, there must exist edges $\{u, s\}, \{v, s\}, \{r, s\} \in E$ (where $u, v, r \in V$ are the vertices whose coordinates have already been calculated), in order to generate a solvable quadratic system

$$\begin{aligned} \|x_s - x_u\|^2 &= d_{us}^2, \\ \|x_s - x_v\|^2 &= d_{vs}^2, \\ \|x_s - x_r\|^2 &= d_{rs}^2, \end{aligned} \tag{3.3}$$

with $x_s \in \mathbb{R}^3$ as the unknown.

b Clearly, in order to guarantee that the system (3.3) will have at most two solutions, the matrix obtained by subtracting one equation from the other two must have full rank.

Exercise 3.7 What geometric interpretation can be given to the assertion that the matrix

$$\begin{bmatrix} x_{v1} - x_{u1} & x_{v2} - x_{u2} & x_{v3} - x_{u3} \\ x_{r1} - x_{u1} & x_{r2} - x_{u2} & x_{r3} - x_{u3} \end{bmatrix}$$

has full rank?

Exercise 3.8 In $\mathbb{R}^2$, for the above exercise, how would the question be answered?

Exercise 3.9 Which is the most important condition in practice—that there exist edges $\{u, s\}, \{v, s\}, \{r, s\} \in E$ or that the associate matrix has complete rank?

Exercise 3.10 If we have more than three edges in conditions (a) and (b) above for Eq. (3.3), how can we choose among these edges?

3.3 Vertex Order for the DGP

Based on the previous discussions, we highlight two important points related to vertices $s \in V$ whose coordinates still need to be positioned in $\mathbb{R}^3$:

1. There are $u, v, r \in V$ such that $\{u, s\}, \{v, s\}, \{r, s\} \in E$,
2. $x_u, x_v, x_r \in \mathbb{R}^3$ are part of a valid realization.

The idea which is the link between these two points is related to the concept of *order* on the vertices of the DGP graph. That is, if there exists an order relation in V which satisfies the conditions 1 and 2 above, we can ensure (excluding rotations, translations, and reflections and supposing that the points related to the vertices u, v, r are not collinear) that the DGP solution set is finite.

Recall that to solve a DGP, for $K = 3$, we must obtain a valid realization $x : V \to \mathbb{R}^3$ of the related graph, which implies that we need to find the coordinates of the points $x_s \in \mathbb{R}^3$ for each $s \in V$, satisfying all equations of the system

$$\forall \{u, v\} \in E, \ \|x_u - x_v\| = d_{uv}.$$

A solution of the problem can then be represented as an element of $\mathbb{R}^{3|V|}$.

Exercise 3.11 If we have such an order on the vertices V, does the search space change?

If three points are given (satisfying the DGP equations) as positions for the first three vertices of the order we are looking for (which is easy to find for most applications), we can say in polynomial time whether or not there is such an order [28, 45].

Consider a DGP with $V = \{a, b, c, u, v, r\}$ and

$$E = \{\{a, b\}, \{a, c\}, \{a, u\}, \{b, c\}, \{b, v\}, \{c, r\}, \{u, v\}, \{u, r\}, \{v, r\}\}.$$

For $K = 2$, we wish to find an order such that the two first vertices generate a clique and, from the third on, there are two previous vertices coming before it. By considering all possible initial cliques, let us see what happens:

- *Starting with the clique $\{a, b\}$, we have the following possible vertices for the third position: c, u, v, r. The next vertex would be c, because $\{\{a, c\}, \{b, c\}\} \subset E$. However, after that, there would exist no other candidate vertices.*
- *Starting with the clique $\{a, c\}$, we have the following possible vertices for the third position: b, u, v, r. The next vertex would be b, because $\{\{a, b\}, \{c, b\}\} \subset E$. However, after that, there would exist no other candidate vertices.*
- *Starting with the clique $\{a, u\}$, there does not exist any candidate vertex to occupy the third position.*
- *Starting with the clique $\{b, c\}$, we have the following possible vertices for the third position: a, u, v, r. The next vertex would be a, because $\{\{b, a\}, \{c, a\}\} \subset E$. However, after that, there would exist no other candidate vertices.*
- *Starting with the clique $\{b, v\}$, there does not exist any candidate vertex to occupy the third position.*
- *Starting with the clique $\{c, r\}$, there does not exist any candidate vertex to occupy the third position.*
- *Starting with the clique $\{u, v\}$, we have the following possible vertices for the third position: a, b, c, r. The next vertex would be r, because $\{\{u, r\}, \{v, r\}\} \subset E$. However, after that, there would exist no other candidate vertices.*
- *Starting with the clique $\{u, r\}$, we have the following possible vertices for the third position: a, b, c, v. The next vertex would be v, because $\{\{u, v\}, \{r, v\}\} \subset E$. However, after that, there would exist no other candidate vertices.*

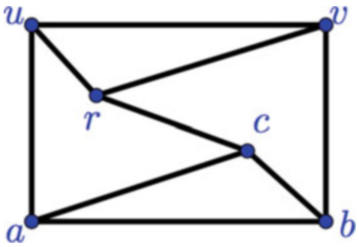


Fig. 3.3 DGP graph without order satisfying condition 1, for K = 2

- *Starting with the clique* $\{v, r\}$, *we have the following possible vertices for the third position:* a, b, c, u. *The next vertex would be* u, *because* $\{\{v, u\}, \{r, u\}\} \subset E$. *However, after that, there would exist no other candidate vertices.*

Therefore, no such order exists! However, the associated graph does not have an uncountable number of realizations (modulo rotations, translations, and reflections) [37] (see Fig. 3.3). That is, the existence of a vertex order mentioned before is just a sufficient condition for the finiteness of the DGP solution set. This question is related to another area of research called *graph rigidity* [32].

Chapter 4
The Discretizable Distance Geometry Problem (DDGP$_3$)

4.1 Definition of the DDGP$_3$

We begin this chapter by describing an important class of the DGP in $\mathbb{R}^3$ having a vertex order as described in Sect. 3.3, called the Discretizable DGP$_3$ (DDGP$_3$). Even though this definition can be extended to $\mathbb{R}^K$ [65], we will consider just the case $K = 3$.

Definition 4.1 (DDGP$_3$) Given a graph $G = (V, E, d)$ of a DGP with $K = 3$ and an order on the vertices V, denoted by $v_1, \ldots, v_n$, such that:

- *there is a valid realization for* v_1, v_2, v_3*,*
- *for all* v_i*,* $i = 4, \ldots, n$*, there are (at least) three previous vertices* a_i, b_i, c_i *with* $\{\{a_i, v_i\}, \{b_i, v_i\}, \{c_i, v_i\}\} \subset E$ satisfying

$$d_{a_i b_i} + d_{b_i c_i} > d_{a_i c_i},$$

find a function $x : V \to \mathbb{R}^3$ *such that*

$$\forall \{v_i, v_j\} \in E, \quad \|x_{v_i} - x_{v_j}\| = d_{v_i v_j}.$$

The DDGP$_3$ is obviously a particular case of the DGP, where $K = 3$, with an order structure given as part of the problem.

The order structure on the vertex set of the associated graph of a DDGP$_3$ allows us to attack the problem taking into account this information. Orders on the vertices of a graph appear in many applications [12, 31]. Essentially, the idea is to proceed vertex by vertex, following the given order.

Let us consider a DGP with $K = 3$, given by

$$V = \{p, q, r, s, t, u, v\}$$

C. Lavor et al., *An Introduction to Distance Geometry applied to Molecular Geometry*, SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-57183-6_4

and

$$\begin{aligned} E = \{&\{p,q\},\{p,r\},\{p,s\},\{p,u\},\{p,v\},\\ &\{q,r\},\{q,s\},\{q,t\},\{q,u\},\{q,v\},\\ &\{r,s\},\{r,t\},\{r,v\},\\ &\{s,t\},\{s,v\},\\ &\{t,u\},\{t,v\},\\ &\{u,v\}\}. \end{aligned}$$

We can first order the vertices according to the process described in Sect. 3.3, resulting in the following:

$$V = \{r,q,t,s,v,u,p\}.$$

Given this order, we have a DDGP$_3$ (supposing that the related hypothesis are satisfied) with

$$\begin{aligned} \{\{r,s\},\{q,s\},\{t,s\}\} &\subset E,\\ \{\{r,v\},\{q,v\},\{t,v\},\{s,v\}\} &\subset E,\\ \{\{q,u\},\{t,u\},\{v,u\}\} &\subset E,\\ \{\{r,p\},\{q,p\},\{s,p\},\{v,p\},\{u,p\}\} &\subset E. \end{aligned}$$

Rewriting the vertices in V as $r = v_1$, $q = v_2$, $t = v_3$, $s = v_4$, $v = v_5$ $u = v_6$, and $p = v_7$, the associated graph for our example is given by

$$V = \{v_1,v_2,v_3,v_4,v_5,v_6,v_7\}$$

and

$$\begin{aligned} E = \{&\{v_1,v_2\},\{v_1,v_3\},\{v_1,v_4\},\{v_1,v_5\},\{v_1,v_7\},\\ &\{v_2,v_3\},\{v_2,v_4\},\{v_2,v_5\},\{v_2,v_6\},\{v_2,v_7\},\\ &\{v_3,v_4\},\{v_3,v_5\},\{v_3,v_6\},\\ &\{v_4,v_5\},\{v_4,v_7\},\\ &\{v_5,v_6\},\{v_5,v_7\},\\ &\{v_6,v_7\}\}. \end{aligned}$$

In order to facilitate the construction of the quadratic systems, let us consider the edges related to v_4, v_5, v_6, v_7:

$$\{\{v_1, v_4\}, \{v_2, v_4\}, \{v_3, v_4\}\} \subset E,$$
$$\{\{v_1, v_5\}, \{v_2, v_5\}, \{v_3, v_5\}, \{v_4, v_5\}\} \subset E,$$
$$\{\{v_2, v_6\}, \{v_3, v_6\}, \{v_5, v_6\}\} \subset E,$$
$$\{\{v_1, v_7\}, \{v_2, v_7\}, \{v_4, v_7\}, \{v_5, v_7\}, \{v_6, v_7\}\} \subset E.$$

From the hypothesis of the DDGP$_3$, v_1, v_2, v_3 are already fixed in $\mathbb{R}^3$, that is, their coordinates $x_{v_1}, x_{v_2}, x_{v_3}$ have already been computed. To obtain the coordinates of v_4, we can exploit $\{\{v_1, v_4\}, \{v_2, v_4\}, \{v_3, v_4\}\} \subset E$, which in turn generates a system of quadratic equations with $x_{v_4} \in \mathbb{R}^3$ as the only unknown variable:

$$\|x_{v_4} - x_{v_1}\|^2 = d^2_{v_1 v_4},$$
$$\|x_{v_4} - x_{v_2}\|^2 = d^2_{v_2 v_4},$$
$$\|x_{v_4} - x_{v_3}\|^2 = d^2_{v_3 v_4}.$$

Expanding the norms and subtracting, for example, the first equation from the other two, we obtain a linear system

$$Ax_{v_4} = b,$$

where $A \in \mathbb{R}^{2\times 3}$ and $b \in \mathbb{R}^2$. Since A has full rank by the hypothesis of the DDGP$_3$ (no collinearities), we obtain x_{v_4} in terms of a parameter $\lambda \in \mathbb{R}$ which we denote by $x_{v_4}(\lambda)$. Replacing $x_{v_4}(\lambda)$ in one of the equations of the quadratic system above, for example, in the equations $\|x_{v_4} - x_{v_1}\|^2 = d^2_{v_1 v_4}$, we obtain the following second-degree equation in λ:

$$\|x_{v_4}(\lambda)\|^2 - 2x_{v_4}(\lambda) \cdot x_{v_1} + \|x_{v_1}\|^2 - d^2_{v_1 v_4} = 0. \tag{4.1}$$

With the values of λ, we obtain the possible results for the coordinates of $x_{v_4}(\lambda)$.

Exercise 4.1 When we obtain two equal roots as a solution to (4.1), what does this means geometrically?

Exercise 4.2 Could we obtain complex roots to (4.1)?

For each solution for v_4 ($x^0_{v_4}$ and $x^1_{v_4}$), performing the same procedure described above, we will have two solutions for v_5. That is, in our example, considering $a_5 = v_2, b_5 = v_3, c_5 = v_4$, and choosing $x^0_{v_4}$, we have the following quadratic system:

$$\|x_{v_5} - x_{v_2}\|^2 = d^2_{v_2 v_5},$$
$$\|x_{v_5} - x_{v_3}\|^2 = d^2_{v_3 v_5},$$
$$\|x_{v_5} - x^0_{v_4}\|^2 = d^2_{v_4 v_5}.$$

However, we also have $\{v_1, v_5\} \in E$, which can be used to test each of the possible positions obtained for v_5 ($x^0_{v_5}$ and $x^1_{v_5}$):

$$\|x^0_{v_5} - x_{v_1}\| = d_{v_1 v_5} \text{ or } \|x^1_{v_5} - x_{v_1}\| = d_{v_1 v_5}?$$

Supposing that the points $\{x_{v_1}, x_{v_2}, x_{v_3}, x^0_{v_4}\}$ are not coplanar, only one of these equations must be satisfied. Suppose it is the first: $x^0_{v_5}$.

Exercise 4.3 What is the difference if we had selected $x^1_{v_4}$ instead of $x^0_{v_4}$?

Exercise 4.4 Why have we assumed the non-coplanarity of the points $\{x_{v_1}, x_{v_2}, x_{v_3}, x^0_{v_4}\}$?

Let us go to the next vertex: v_6. Since we only have $\{\{v_2, v_6\}, \{v_3, v_6\}, \{v_5, v_6\}\} \subset E$, we generate the associated quadratic system

$$\begin{aligned} \|x_{v_6} - x_{v_2}\|^2 &= d^2_{v_2 v_6}, \\ \|x_{v_6} - x_{v_3}\|^2 &= d^2_{v_3 v_6}, \\ \|x_{v_6} - x^0_{v_5}\|^2 &= d^2_{v_5 v_6}, \end{aligned}$$

obtaining the two possible solutions: $x^0_{v_6}$ and $x^1_{v_6}$. At this point, nothing can be said about the feasibility of these points because we do not have more vertices linked to v_6. In order to continue, we have to choose one of the two solutions that were generated, for example, $x^0_{v_6}$. At the same time, we must remember that there is the second possibility: $x^1_{v_6}$.

For the next vertex v_7, considering

$$\begin{aligned} a_7 &= v_4, \\ b_7 &= v_5, \\ c_7 &= v_6, \end{aligned}$$

we have a new quadratic system,

$$\begin{aligned} \|x_{v_7} - x^0_{v_4}\|^2 &= d^2_{v_4 v_7}, \\ \|x_{v_7} - x^0_{v_5}\|^2 &= d^2_{v_5 v_7}, \\ \|x_{v_7} - x^0_{v_6}\|^2 &= d^2_{v_6 v_7}, \end{aligned}$$

obtaining two more possible solutions: $x^0_{v_7}$ and $x^1_{v_7}$. In this case, there are two other edges, $\{\{v_1, v_7\}, \{v_2, v_7\}\} \in E$, that can be used to test each one of the possible positions for v_7:

$$\|x^0_{v_7} - x_{v_1}\| = d_{v_1 v_7} \text{ or } \|x^1_{v_7} - x_{v_1}\| = d_{v_1 v_7}$$

and

$$\|x^0_{v_7} - x_{v_2}\| = d_{v_2 v_7} \text{ or } \|x^1_{v_7} - x_{v_2}\| = d_{v_2 v_7}.$$

Now we have two possibilities:

- Neither of the possible solutions are feasible,
- Only one of the possible solutions is feasible.

The first case can occur if the selection that we made for v_6 had been wrong. In this case, we need to return, choose $x^1_{v_6}$, and repeat the process. In case we have "extra" edges, we must test them, one by one. Just one equation that is not satisfied will cause the process to terminate and require a backtracking.

In the second case of one feasible solution, we have a situation like that for vertex v_5, but with two extra edges instead of only one: $\{v_1, v_7\} \in E$ and $\{v_2, v_7\} \in E$. Theoretically, we can use any of these edges for choosing between $x^0_{v_7}$ and $x^1_{v_7}$.

Let us suppose that neither of the possible solutions are feasible. Then, we "return" , choose $x^1_{v_6}$, and solve a corresponding quadratic system, obtaining $x^0_{v_7}$ and $x^1_{v_7}$. Since we have $\{v_1, v_7\} \in E$ (supposing that the points $\{x_{v_1}, x_{v_2}, x^0_{v_4}, x^1_{v_6}\}$ are not coplanar), only one of these positions is feasible. Suppose it is $x^1_{v_7}$. Therefore, we finally get a solution to the problem: $\{x_{v_1}, x_{v_2}, x_{v_3}, x^0_{v_4}, x^0_{v_5}, x^1_{v_6}, x^1_{v_7}\}$.

Exercise 4.5 When we have to choose between $x^0_{v_6}$ and $x^1_{v_6}$, is it possible to anticipate that $x^0_{v_6}$ will be infeasible?

Exercise 4.6 By using the procedure above, what ensures us that we will find a solution?

Exercise 4.7 In practice, which criteria can be adopted in order to decide which edge must be used if there are many extra edges and supposing that all of them are theoretically "feasible" ?

4.2 Complexity of the DDGP$_3$

The computational cost of the method described above may be exponential. However, is it possible to create an efficient algorithm for the DDGP$_3$? In Sect. 2.3, we mentioned that the DGP is an NP-hard problem. Since the DDGP$_3$ is a subproblem of the DGP, it is natural to ask if there is some change in the computational complexity of the subproblem. However, the DDGP$_3$ remains NP-hard [65].

Recall that in the order associated with the graph $G = (V, E)$ of the DDGP$_3$, for all $v_i \in V,\ i = 4, \dots, n$, there exists at least three previous vertices a_i, b_i, c_i with

$$\{\{a_i, v_i\}, \{b_i, v_i\}, \{c_i, v_i\}\} \subset E.$$

Remark 4.1 In the example above, if we have $\{v_1, v_6\} \in E$ or $\{v_4, v_6\} \in E$, no backtracking is necessary. If we generalize this fact, we can say that, if for all $v_i \in V,\ i = 5, \dots, n$, there exists at least four previous vertices a_i, b_i, c_i, d_i with

$$\{\{a_i, v_i\}, \{b_i, v_i\}, \{c_i, v_i\}, \{d_i, v_i\}\} \subset E,$$

(supposing that $\{a_i, b_i, c_i, d_i\}$ are related to noncoplanar points), the DDGP$_3$ can be solved in linear time!

Exercise 4.8 If we have a DDGP$_3$ whose solution set is not empty and for all $v_i \in V,\ i = 5, \dots, n$, there exists (at least) four previous vertices a_i, b_i, c_i, d_i with

$$\{\{a_i, v_i\}, \{b_i, v_i\}, \{c_i, v_i\}, \{d_i, v_i\}\} \subset E$$

and $\{a_i, b_i, c_i, d_i\}$ are not coplanar, can we assert that there is only one solution?

4.3 The BP Algorithm for the DDGP$_3$

The vertex order in the DDGP$_3$ ensures the finiteness of the solution set (assuming that the respective triangle inequalities are strictly satisfied) and allow us to "organize" the search space appropriately. Although the search space remains continuous ($\mathbb{R}^{3n}$, with n vertices), the vertex order indicates the way in which the search space should be covered so that a solution of the problem can be found. In fact, the vertex order induces a structure of a binary tree in the search space (Fig. 4.1), where the root represents the coordinates of x_{v_1}, and the next two vertices represent the coordinates of x_{v_2} and x_{v_3} that have been fixed. From the fourth level of the associated tree, we have the representations of all the possible positions for the vertices $v_i,\ i = 4, \dots, n$. There are two possibilities for v_4, 4 for v_5, 8 for v_6, 16 for v_7, . . ., 2^{i-3} for v_i, . . ., and 2^{n-3} for v_n.

From what we have seen from the previous example, we decided to explore the tree from the left to the right side. That is, when we do not have additional edges,

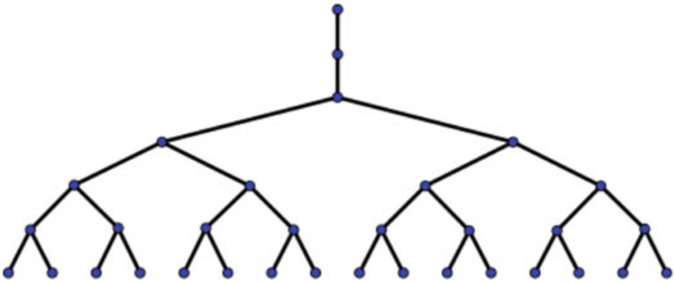


Fig. 4.1 Binary tree

we opted by choosing solution $x^0_{v_i}$ of the associated quadratic system. Other options can be defined [25, 69], but we need to take care in order to not get lost when we have to backtrack the tree.

Based on the previous example, we can develop a procedure to solve the DDGP$_3$ that consists of a sequence (induced by the order on the vertices) of quadratic systems and feasibility tests (when additional edges exist). Thus, we can divide the edges E of the graph of the DDGP$_3$ in two disjoint sets:

$$E = E_d \cup E_p,$$

where

$$E_d = \{\{a_4, v_4\}, \{b_4, v_4\}, \{c_4, v_4\}, \ldots, \{a_n, v_n\}, \{b_n, v_n\}, \{c_n, v_n\}\}$$

is the *branching set* and

$$E_p = E - E_d$$

is the *pruning set*.

The branching edges "model" the search space as a binary tree and the pruning edges "indicate" the way that we should proceed, going down the tree. To find a solution, we go down the tree from the root until we reach the last level, performing all feasibility tests along the way. The solution is given by the path defined only by the feasible nodes.

We saw that, in some cases, we have to backtrack the tree and restart the path. These backtracks can exponentially increase the computational cost of the search. We can traverse "down" the tree without backtracking in two special cases:

- $E_p = \emptyset$,
- $\forall v_i, i = 5, \ldots, n$, there are (at least) four vertices which generate (non-coplanar) previous points to v_i, given by a_i, b_i, c_i, d_i, such that $\{\{a_i, v_i\}, \{b_i, v_i\}, \{c_i, v_i\}, \{d_i, v_i\}\} \subset E$.

In both cases, a solution can be quickly found with a computational cost proportional to n.

For the first case, it is sufficient to choose any of the solutions of the quadratic systems

$$\|x_{v_i} - x_{a_i}\|^2 = d^2_{a_i v_i},$$
$$\|x_{v_i} - x_{b_i}\|^2 = d^2_{b_i v_i},$$
$$\|x_{v_i} - x_{c_i}\|^2 = d^2_{c_i v_i},$$

since there is no feasibility test to do.

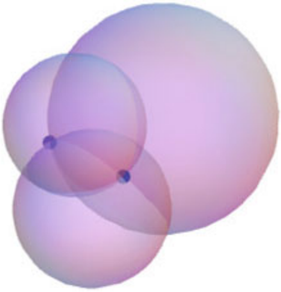


Fig. 4.2 Intersection of three spheres

For the second case, it is sufficient to solve the quadratic systems

$$\|x_{v_i} - x_{a_i}\|^2 = d^2_{a_i v_i},$$
$$\|x_{v_i} - x_{b_i}\|^2 = d^2_{b_i v_i},$$
$$\|x_{v_i} - x_{c_i}\|^2 = d^2_{c_i v_i},$$
$$\|x_{v_i} - x_{d_i}\|^2 = d^2_{d_i v_i},$$

which provides a unique solution. There is also no feasibility tests, since we included them in the quadratic systems.

The set E_d is essentially the "same" for any DDGP$_3$, whereas the set E_p depends on the particular problem. The branching edges define the quadratic systems and the pruning edges define the feasibility tests.

Consider a generic quadratic system associated with $v_i, i = 4, \ldots, n$:

$$\|x_{v_i} - x_{a_i}\|^2 = d^2_{a_i v_i},$$
$$\|x_{v_i} - x_{b_i}\|^2 = d^2_{b_i v_i},$$
$$\|x_{v_i} - x_{c_i}\|^2 = d^2_{c_i v_i}.$$

Geometrically, to solve this system means to obtain the intersection of three spheres (Fig. 4.2), $S(x_{a_i}, d_{a_i v_i}), S(x_{b_i}, d_{b_i v_i}), (S_{c_i}, d_{c_i v_i})$, centered at $x_{a_i}, x_{b_i}, x_{c_i} \in \mathbb{R}^3$ with radius $d_{a_i v_i}, d_{b_i v_i}, d_{c_i v_i} \in (0, \infty)$, given by

$$D = S(x_{a_i}, d_{a_i v_i}) \cap S(x_{b_i}, d_{b_i v_i}) \cap (S_{c_i}, d_{c_i v_i}).$$

From the hypothesis of the strict triangle inequality of the DDGP$_3$, the points a_i, b_i, c_i are not collinear and, therefore, we only have three possibilities:

- $|D| = 0$,
- $|D| = 1$,
- $|D| = 2$.

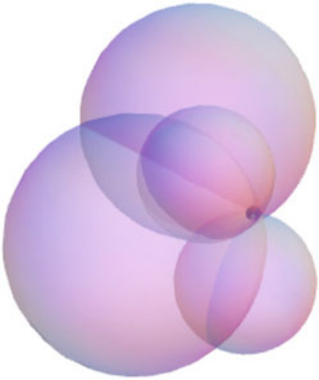


Fig. 4.3 Intersection of four spheres

When $E_p \neq \emptyset$, we have additional spheres (Fig. 4.3) that should be added to the previous intersection. That is, we should have

$$P = D \cap \left(\cap_{i=1}^{k_i} S(x_{p_i}, d_{p_i v_i}) \right),$$

where k_i is the number of additional edges to $v_i, p_i < v_i, \{p_i, v_i\} \in E$, and $S(x_{p_i}, d_{p_i v_i})$ is the sphere with center at point x_{p_i} with radius $d_{p_i v_i}$. This new intersection, $\cap_{i=1}^{k_i} S(x_{p_i}, d_{p_i v_i})$, is precisely the feasibility test. With additional spheres, we have only two possibilities:

- $|P| = 0$,
- $|P| = 1$.

The first case corresponds to some infeasibility, and the second one corresponds to the case where we have the point satisfying all the given distances to previous points already positioned. This geometrical interpretation gives us other alternatives to deal with the two fundamental subproblems that are involved in solution methods for the DDGP$_3$ (see [4–6, 49, 67]). The algorithm to find a solution for the DDGP$_3$ is called *Branch and Prune* [44, 53], or just *BP*.

Exercise 4.9 Write down the main steps of the BP algorithm as a pseudo code.

Exercise 4.10 Explain that, when the points $x_{a_i}, x_{b_i}, x_{c_i}, x_{v_i}$ are coplanar, we have the case $|D| = 1$.

Exercise 4.11 When $E_p = \emptyset$, is it possible to know how many solutions the problem has?

Exercise 4.12 Does the BP find all solutions of a DDGP$_3$?

Chapter 5
The Discretizable Molecular Distance Geometry Problem (DMDGP)

5.1 Definition of the DMDGP

We know that to ensure the finiteness of the solution set of the DGP, we can impose an order on the vertices of the associated graph. If such an order exists, it is not hard to find it in the DGP graph.

The DDGP$_3$ assumes that, for all v_i, $i = 4, \ldots, n$, there exist (at least) three previous vertices a_i, b_i, c_i with $\{\{a_i, v_i\}, \{b_i, v_i\}, \{c_i, v_i\}\} \subset E$, such that

$$d_{a_i b_i} + d_{b_i c_i} > d_{a_i c_i}. \tag{5.1}$$

Depending on the DDGP$_3$ instance, some distances between the vertices a_i, b_i, c_i may be lacking, which may imply no solution in $\mathbb{R}^3$ for the quadratic system

$$\begin{aligned} \|x_{v_i} - x_{a_i}\|^2 &= d^2_{a_i v_i}, \\ \|x_{v_i} - x_{b_i}\|^2 &= d^2_{b_i v_i}, \\ \|x_{v_i} - x_{c_i}\|^2 &= d^2_{c_i v_i}. \end{aligned}$$

A way to avoid this mishap is to require that, for all v_i, $i = 4, \ldots, n$, the distances between the vertices a_i, b_i, c_i are known (note that this is not a requirement in the Definition of the DDGP$_3$). Additionally, we may require that the vertices a_i, b_i, c_i are the immediate predecessors to v_i, which occurs in many applications [57].

Exercise 5.1 Geometrically, what does it mean for the quadratic system above to have no solution?

Exercise 5.2 If the vertices a_i, b_i, c_i compose a clique, can we ensure that the related triangle inequality is strictly satisfied?

C. Lavor et al., *An Introduction to Distance Geometry applied to Molecular Geometry*, SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-57183-6_5

We are now interested in finding a vertex order in which the vertices used in the construction of each quadratic system compose a clique and are the immediate predecessors to the vertex whose coordinates we wish to calculate. Considering the same DGP instance as in Sect. 4.1, does there exist an order with these properties? Let us study this question before proceeding.

Let us return to that problem: a DGP with $K = 3$, where the associated graph is $G = (V, E)$, with vertices $V = \{p, q, r, s, t, u, v\}$ and edges

$$\begin{aligned} E = \{&\{p,q\},\{p,r\},\{p,s\},\{p,u\},\{p,v\},\\ &\{q,r\},\{q,s\},\{q,t\},\{q,u\},\{q,v\},\\ &\{r,s\},\{r,t\},\{r,v\},\\ &\{s,t\},\{s,v\},\\ &\{t,u\},\{t,v\},\\ &\{v,u\}\}. \end{aligned}$$

Consider a new ordering given by

$$V = \{p, u, v, q, t, r, s\}.$$

We use the following notation in order to facilitate our analysis: $p = u_1$, $u = u_2$, $v = u_3$, $q = u_4$, $t = u_5$, $r = u_6$, and $s = u_7$ (u_i is used instead of v_i in order to emphasize that we are using a different order from the previous one). We can verify that this order has the desired properties, because in addition to the valid realization for $\{u_1, u_2, u_3\}$, we also have the following cliques:

$$\begin{aligned} &\{u_1, u_2, u_3, u_4\},\\ &\{u_2, u_3, u_4, u_5\},\\ &\{u_3, u_4, u_5, u_6\},\\ &\{u_4, u_5, u_6, u_7\}. \end{aligned}$$

To better follow the BP algorithm, we note that

$$\begin{aligned} \{\{u_1,u_4\},\{u_2,u_4\},\{u_3,u_4\}\} &\subset E,\\ \{\{u_2,u_5\},\{u_3,u_5\},\{u_4,u_5\}\} &\subset E,\\ \{\{u_1,u_6\},\{u_3,u_6\},\{u_4,u_6\},\{u_5,u_6\}\} &\subset E,\\ \{\{u_1,u_7\},\{u_3,u_7\},\{u_4,u_7\},\{u_5,u_7\},\{u_6,u_7\}\} &\subset E. \end{aligned}$$

To obtain the coordinates of u_4, we consider the following quadratic system, with $x_{u_4} \in \mathbb{R}^3$ as the only variable:

$$\begin{aligned}
||x_{u_4} - x_{u_1}||^2 &= d^2_{u_1u_4},\\
||x_{u_4} - x_{u_2}||^2 &= d^2_{u_2u_4},\\
||x_{u_4} - x_{u_3}||^2 &= d^2_{u_3u_4}.
\end{aligned}$$

We choose one of the two possible values for the coordinates of u_4, let us say $x^0_{u_4}$, and we obtain the following quadratic system in order to find the coordinates of u_5,

$$\begin{aligned}
||x_{u_5} - x_{u_2}||^2 &= d^2_{u_2u_5},\\
||x_{u_5} - x_{u_3}||^2 &= d^2_{u_3u_5},\\
||x_{u_5} - x^0_{u_4}||^2 &= d^2_{u_4u_5}.
\end{aligned}$$

Again, we choose one of the two possible values for the coordinates of u_5, let us say $x^0_{u_5}$, and we obtain a new quadratic system,

$$\begin{aligned}
||x_{u_6} - x_{u_3}||^2 &= d^2_{u_3u_6},\\
||x_{u_6} - x^0_{u_4}||^2 &= d^2_{u_4u_6},\\
||x_{u_6} - x^0_{u_5}||^2 &= d^2_{u_5u_6}.
\end{aligned}$$

We also have $\{u_1, u_6\} \in E$, that we can use to check which one of the possible coordinates obtained for u_6, $x^0_{u_6}$ and $x^1_{u_6}$, is feasible:

$$||x^0_{u_6} - x_{u_1}|| = d_{u_1u_6} \quad \text{or} \quad ||x^1_{u_6} - x_{u_1}|| = d_{u_1u_6}?$$

Maybe none of the equations is satisfied, implying that we made a wrong choice for u_5. Let us suppose that it is the case. We need to return and recompute the coordinates using $x^1_{u_5}$. The new quadratic system is

$$\begin{aligned}
||x_{u_6} - x_{u_3}||^2 &= d^2_{u_3u_6},\\
||x_{u_6} - x^0_{u_4}||^2 &= d^2_{u_4u_6},\\
||x_{u_6} - x^1_{u_5}||^2 &= d^2_{u_5u_6}.
\end{aligned}$$

Again, by using $\{u_1, u_6\} \in E$, we check each of the new possible positions obtained for u_6 ($y^0_{u_6}$ and $y^1_{u_6}$):

$$||y^0_{u_6} - x_{u_1}|| = d_{u_1u_6} \quad \text{or} \quad ||y^1_{u_6} - x_{u_1}|| = d_{u_1u_6}?$$

Supposing that the points $\{x_{u_1}, x_{u_3}, x^0_{u_4}, x^1_{u_5}\}$ are not coplanar, only one of these equations will be satisfied. Let us suppose that it is the first. Then, we discard $y^1_{u_6}$ and we consider $y^0_{u_6}$. The new quadratic system is

$$\|x_{u_7} - x^0_{u_4}\|^2 = d^2_{u_4u_7},$$
$$\|x_{u_7} - x^1_{u_5}\|^2 = d^2_{u_5u_7},$$
$$\|x_{u_7} - y^0_{u_6}\|^2 = d^2_{u_6u_7}.$$

By using $\{u_1, u_7\} \in E$ and $\{u_3, u_7\} \in E$ to check each of the new possible positions obtained for u_7 ($x^0_{u_7}$ and $x^1_{u_7}$), we have:

$$\|x^0_{u_7} - x_{u_1}\| = d_{u_1u_7} \quad \text{or} \quad \|x^1_{u_7} - x_{u_1}\| = d_{u_1u_7}$$

and

$$\|x^0_{u_7} - x_{u_3}\| = d_{u_3u_7} \quad \text{or} \quad \|x^1_{u_7} - x_{u_3}\| = d_{u_3u_7}.$$

Supposing that $x^1_{u_7}$ is selected, we therefore obtain the solution to our problem as

$$x_{u_1}, x_{u_2}, x_{u_3}, x^0_{u_4}, x^1_{u_5}, y^0_{u_6}, x^1_{u_7}.$$

Until this moment, it was not possible to see any advantage in this new order, besides the fact that it ensures that the quadratic systems have solutions. However, with the cliques $\{v_{i-3}, v_{i-2}, v_{i-1}, v_i\}$, for all $i = 4, \ldots, n$, we can replace the solution of the quadratic systems by something numerically simpler and more stable. Before proceeding with explaining how to do this, we will formalize the definition of the new problem: the *Discretizable Molecular Distance Geometry Problem* (DMDGP).

Definition 5.1 (DMDGP) Given a graph $G = (V, E, d)$ of a *DGP* with $K = 3$ and an order on the vertices V, denoted by $v_1, \ldots, v_n$, such that

- there is a valid realization for v_1, v_2, v_3,
- for all v_i, $i = 4, \ldots, n$, there are (at least) three immediately previous vertices $v_{i-3}, v_{i-2}, v_{i-1}$, where $\{v_{i-3}, v_{i-2}, v_{i-1}, v_i\}$ is a clique, and

$$d_{v_{i-3}v_{i-2}} + d_{v_{i-2}v_{i-1}} > d_{v_{i-3}v_{i-1}},$$

find a function $x : V \to \mathbb{R}^3$ such that

$$\forall \{v_i, v_j\} \in E, \ \|x_{v_i} - x_{v_j}\| = d_{v_iv_j}.$$

5.2 Complexity of the DMDGP

The existence of the cliques $\{v_{i-3}, v_{i-2}, v_{i-1}, v_i\}$, for all $i = 4, \ldots, n$, provides us more information about the vertex order of the associated DGP graph. By using the distance information of the cliques $\{v_{i-3}, v_{i-2}, v_{i-1}, v_i\}$, we are able to obtain "almost" all of the following values:

- $d_{1,2}, \dots, d_{n-1,n}$: distances associated with the consecutive vertices,
- $\theta_{1,3}, \dots, \theta_{n-2,n}$: planar angles associated with three consecutive vertices,
- $\omega_{1,4}, \dots, \omega_{n-3,n}$: torsion angles associated with four consecutive vertices.

Exercise 5.3 What is the reason of the "almost" above?

Recall that the torsion angle $\omega_{i-3,i}$ is the angle between the normal vectors associated with the planes determined by the vertices $v_{i-3}, v_{i-2}, v_{i-1}$ and v_{i-2}, v_{i-1}, v_i, respectively. The values $d_{1,2}, \dots, d_{n-1,n}$ are, obviously, obtained from the definition of the DMDGP, and the values $\theta_{1,3}, \dots, \theta_{n-2,n}$ are obtained by the law of cosines. However, from the DMDGP hypothesis, we can obtain only the values of the cosines of the torsion angles, given by $(i = 4, \dots, n)$ [44, 49]:

$$\cos(\omega_{i-3,i}) = \frac{2d^2_{i-2,i-1}(d^2_{i-3,i-2} + d^2_{i-2,i} - d^2_{i-3,i}) - (d_{i-3,i-2,i-1})(d_{i-2,i-1,i})}{\sqrt{4d^2_{i-3,i-2}d^2_{i-2,i-1} - (d^2_{i-3,i-2,i-1})}\sqrt{4d^2_{i-2,i-1}d^2_{i-2,i} - (d^2_{i-2,i-1,i})}},$$

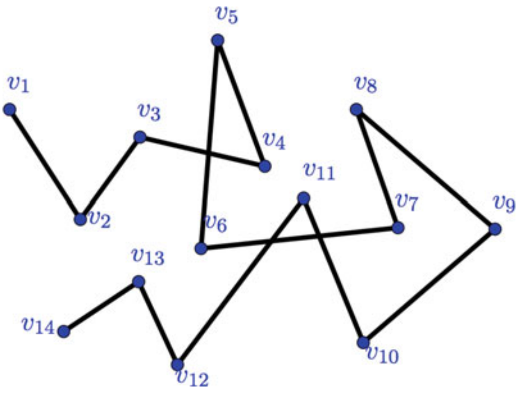
where

$$d_{i-3,i-2,i-1} = d^2_{i-3,i-2} + d^2_{i-2,i-1} - d^2_{i-3,i-1},$$
$$d_{i-2,i-1,i} = d^2_{i-2,i-1} + d^2_{i-2,i} - d^2_{i-1,i}.$$

Actually, with the ordering structure of the DMDGP, we can imagine that we have molecular structures (Fig. 5.1), which explains the use of the extra term "molecular" to describe this new class of problems.

Exercise 5.4 What is the distance that appears only one time in the formula above for $cos(\omega_{i-3,i})$? Why is it "different" from the other distances?

Fig. 5.1 DMDGP instance as a molecule



Remember that to define a three dimensional structure of a molecule, we can use the internal coordinates, given by the values $d_{1,2}, \ldots, d_{n-1,n}$, $\theta_{1,3}, \ldots, \theta_{n-2,n}$, and $\omega_{1,4}, \ldots, \omega_{n-3,n}$. However, as we mentioned above, in the DMDGP, we just have $\cos(\omega_{i-3,i})$, $i = 4, \ldots, n$, which generates two possible values for each torsion angle, since $\omega_{i-3,i} \in [0, 2\pi]$. This implies that we do not need to solve anymore quadratic systems! Moreover, the two possibilities for each torsion angle can be found by using the distances related to the cliques $\{v_{i-3}, v_{i-2}, v_{i-1}, v_i\}$. In order to prune using the given extra distances, we use the matrices given in Sect. 2.4 to obtain the Cartesian coordinates of the two possible solutions for each vertex v_i, where we have already obtained the coordinates of $v_{i-3}, v_{i-2}, v_{i-1}$, and then we just compare the distances we calculate with the given distances.

A question remains: "What is the computational cost in finding the DMDGP order?" This is different than the polynomial cost of obtaining the DDGP$_3$ order. In fact, finding a DMDGP order may be difficult because it is an NP-hard problem [12]. It is the cost we pay for the new information. We escaped solving quadratic systems, but we exponentially increased the cost of finding a DGP order. However, depending on the application, that order can be obtained using the characteristics of the particular problem. It is what happens, for example, in problems related to 3D protein structures [48] (see Chap. 6).

Exercise 5.5 Why is there a "change of signs" in the formulas $d_{i-3,i-2,i-1} = d^2_{i-3,i-2} + d^2_{i-2,i-1} - d^2_{i-3,i-1}$ and $d_{i-2,i-1,i} = d^2_{i-2,i-1} + d^2_{i-2,i} - d^2_{i-1,i}$ above?

Exercise 5.6 Derive the formula for $cos(\omega_{i-3,i})$.

5.3 DMDGP Symmetry

We saw that the DMDGP order allows us to view the problem as a molecule with a finite possible configurations, and by using the internal coordinates and the distance information of the cliques $\{v_{i-3}, v_{i-2}, v_{i-1}, v_i\}$, $i = 4, \ldots, n$, we can also find the two possible values for all torsion angles $\omega_{i-3,i}$. There exists another interesting property related to the symmetry of the DMDGP solutions.

In our example problem from Sect. 5.1, we realized that the two positions for v_4 $(x^0_{v_4}, x^1_{v_4})$ can be considered since there are no extra edges which can invalidate one of them. This implies that, for any solution found in the left subtree, having the node $x^0_{v_4}$ as its root, there exists another one that is symmetric to the plane defined by $x_{v_1}, x_{v_2}, x_{v_3}$ [47]. Note the relation that exists among the finiteness of the solution set, the strict triangle inequality, and the symmetries.

An immediate consequence of this fact is that the solution set has an even number of solutions. However, since the first computational results obtained for the DMDGP [53], we empirically observed that the number of solutions was always a power of two. Only recently, by using group theory, a mathematical proof of this fact was presented [58].

We illustrate the importance of this result by considering the following set, given a DMDGP graph $G = (V, E, d)$ with the vertex order $v_1, \ldots, v_n$:

$$S = \{v \in V : \nexists\{u, w\} \in E \ \text{ such that } u + 3 < v \leq w\}.$$

In order to simplify the notation, we denote by $u + 3$ the third vertex after u, and by $u - 3$ the third vertex before u. Initially, let us try to identify the elements in S. The first candidate is v_4, which is in S if there is no edge $\{u, w\} \in E$ such that $u + 3 < v_4 \leq w$. If there is some $u \in V$ satisfying this property, we will have $u < v_4 - 3$. However, this is not possible because $v_4 - 3 = v_1$, which is the first element of V. That is, $v_4 \in S$ for any DMDGP.

Let us see what happens with v_5 (we are supposing that the DMDGP has a solution):

- Supposing that there exists $\{u, v_5\} \in E$, such that $u < v_5 - 3$, we have that $v_5 \notin S$ and $\{v_1, v_5\} \in E$, which implies that only one of the possibilities for v_5 is feasible: either $x^0_{v_5}$ or $x^1_{v_5}$.
- Supposing that there is no $\{u, v_5\} \in E$, for $u < v_5 - 3$, we need to consider the two following cases:
  - If there is no $\{u, w\} \in E$, such that $u + 3 < v_5 < w$, then $v_5 \in S$.
  - If there exits $\{u, w\} \in E$, such that $u + 3 < v_5 < w$, then $v_5 \notin S$.

Since the procedure above can be applied to all elements of V, we can obtain the set S by using just the DMDGP data, even before we apply BP to solve the problem. But what is the importance of the set S?

The set S identifies other symmetric planes for the DMDGP, in addition to the plane associated with the vertices $\{v_1, v_2, v_3\}$, defined for all DMDGP instances [58].

For example, if $v_5 \in S$, this implies that the two positions for v_5 are feasible, $x^0_{v_5}$ and $x^1_{v_5}$. At the same time, $x^0_{v_5}$ and $x^1_{v_5}$ are part of two different DMDGP solutions [66].

Considering the example problem of Sect. 5.1 and using the notation v_i, we have:

$$\begin{aligned}
V = {} & \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\},\\
E = {} & \{\{v_1, v_2\}, \{v_1, v_3\}, \{v_1, v_4\}, \{v_1, v_6\}, \{v_1, v_7\},\\
& \{v_2, v_3\}, \{v_2, v_4\}, \{v_2, v_5\},\\
& \{v_3, v_4\}, \{v_3, v_5\}, \{v_3, v_6\}, \{v_3, v_7\},\\
& \{v_4, v_5\}, \{v_4, v_6\}, \{v_4, v_7\},\\
& \{v_5, v_6\}, \{v_5, v_7\}, \{v_6, v_7\}\}.
\end{aligned}$$

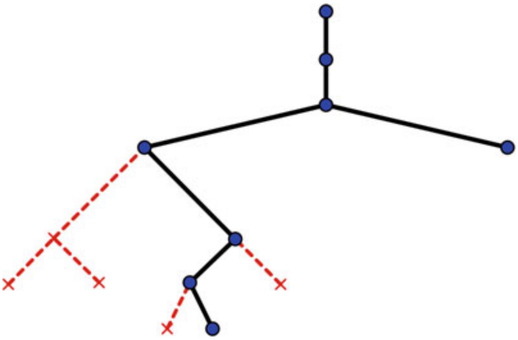


Fig. 5.2 Solution obtained by BP algorithm

It is easy to see that $S = \{v_4\}$, since $\{v_1, v_7\} \in E$. That is, there exists only one symmetric plane (defined by $x_{v_1}, x_{v_2}, x_{v_3}$), which implies that we have only two solutions. As we have already obtained a solution, given by

$$x_{v_1}, x_{v_2}, x_{v_3}, x^0_{v_4}, x^1_{v_5}, y^0_{v_6}, x^1_{v_7},$$

we have another one symmetric to the plane defined by $\{x_{v_1}, x_{v_2}, x_{v_3}\}$ (see Fig. 5.2).

Suppose now that we have a little different DMDGP instance, given by

$$\begin{aligned} V &= \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}, \\ E &= \{\{v_1, v_2\}, \{v_1, v_3\}, \{v_1, v_4\}, \{v_1, v_6\}, \\ &\quad \{v_2, v_3\}, \{v_2, v_4\}, \{v_2, v_5\}, \\ &\quad \{v_3, v_4\}, \{v_3, v_5\}, \{v_3, v_6\}, \\ &\quad \{v_4, v_5\}, \{v_4, v_6\}, \{v_4, v_7\}, \\ &\quad \{v_5, v_6\}, \{v_5, v_7\}, \{v_6, v_7\}\}. \end{aligned}$$

Performing the calculations, we obtain

$$S = \{v_4, v_7\},$$

implying that we have another symmetric plane defined by $\{v_4, v_5, v_6\}$ (see Fig. 5.3).

To simplify the notation, let us represent the first solution by a sequence of zeros and ones and denote the first tree positions by 0, 0, 0:

$$s_1 = (0, 0, 0, 0, 1, 0, 1).$$

Since we know that we have a symmetry at vertex v_7, another solution is given by

$$s_2 = (0, 0, 0, 0, 1, 0, 0).$$

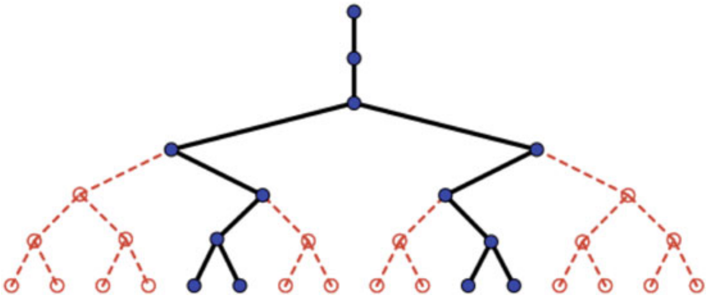


Fig. 5.3 Symmetric solutions

Now, by considering the symmetry at vertex v_4, we obtain other two solutions given by

$$s_3 = (0, 0, 0, 1, 0, 1, 0)$$

and

$$s_4 = (0, 0, 0, 1, 0, 1, 1).$$

We have two important conclusions arising from these observations [1, 55, 66]:

- We know, a priori, using only the data given by any DMDGP, that the cardinality of the solution set is $2^{|S|}$.
- In order to find all the solutions of a DMDGP, it is enough to apply the BP algorithm to find only one solution, since all the others can be obtained using the DMDGP symmetries.

Exercise 5.7 What is the computational importance of knowing a priori the number of DMDGP solutions?

Exercise 5.8 Since the computational cost associated with the use of symmetries to obtain other DMDGP solutions is polynomial, what is the implication for the complexity of DMDGP?

Chapter 6
Distance Geometry and Molecular Geometry

6.1 The DMDGP and 3D Protein Structures

Currently, the most prominent application of distance geometry is related to molecular geometry. Specifically, the problem is the calculation of the 3D protein structure using distance information obtained from Nuclear Magnetic Resonance (NMR) experiments [79, 80]. It is worth mentioning that the 2002 Nobel Prize in Chemistry was awarded to the chemist Kurt Wüthrich for the development of the application of NMR to determine protein structures using distance information related to atoms that are close enough to be detected by NMR experiments.

Why is it important to know the three dimensional structure of a protein molecule? It is because the 3D structure of a molecule is strongly connected with its physicochemical properties. A classical example that illustrates this fact is the discovery of the three dimensional structure of DNA [78]. In 1953, the physicist Maurice Wilkins and the chemist Rosalind Franklin used X-ray diffraction, another technique to determine the structure of proteins [11], to "photograph" the DNA. The problem was to formulate a three dimensional model of a DNA molecule which matched the results of the X-ray diffraction and to explain some known chemical properties. In the same year, the biochemist James Watson and the biophysicist Francis Crick proposed a three dimensional model, the famous double helix, that explained all the available data about the DNA molecule known at the time. The model that arose suggested the mechanism by which transmission of the genetic information was achieved. The essential characteristic of the model is the complementarity of the two twisted strands of DNA. Watson and Crick realized, before the existence of data that verified their model, that the proposed structure could be reproduced by the separation of the two strands and by the synthesis of a complementary strand for each one. In 1958, the molecular biologist Matthew Meselson and the geneticist Franklin Stahl showed experimentally that the Watson and Crick's model of replication of DNA works. With the model and

C. Lavor et al., *An Introduction to Distance Geometry applied to Molecular Geometry*, SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-57183-6_6

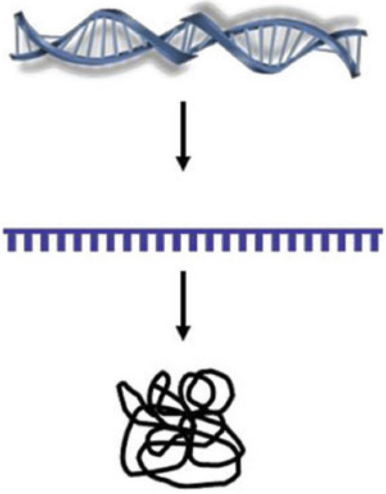


Fig. 6.1 DNA and protein

its experimental verification, a revolution in the understanding of the process of heredity was started. Because of the discovery of the three dimensional structure of the DNA molecule, James Watson shared the 1962 Nobel Prize in Medicine with Francis Crick and Maurice Wilkins.

The genes of a living organism present in DNA are, indirectly, responsible for the physical characteristics of the organism, but the corresponding proteins are, in fact, what determine these characteristics. Inside of the cell, the DNA of a gene is transcribed in the messenger RNA and this transcription is translated in order to form the sequence of amino acids that gives arise to a protein molecule (Fig. 6.1). This process of transcription and translation is well understood [73]. However, there is still much to learn about the mechanism of the formation of the protein molecule from the sequence of amino acids provided by the messenger RNA. This process is called *protein folding* and the associated problem is known as the *protein folding problem* [17].

We have already seen that the determination of the three dimensional structure of a protein molecule is an important problem, but what is the relation to the DMDGP? Havel and Wüthrich, in 1984 and 1985 [35, 36], wrote two articles showing how Distance Geometry can be applied to the calculation of protein structure by using NMR data. However, it was just in 1988 that the book “Distance Geometry and Molecular Conformation” [15] was published. Crippen and Havel established the fundamentals and connections between the two topics of research. Their proposed algorithm, called EMBED, uses the methods of linear algebra and optimization to solve the associated DGP.

Our proposal is to consider the problem as a DMDGP. For this, it is necessary to define an order on the atoms of a protein molecule which induces a vertex order on the corresponding DMDGP graph, given by $v_1, \ldots, v_n$. That is, we must have a valid realization for v_1, v_2, v_3 and, for all v_i, $i = 4, \ldots, n$, there must exist three immediate previous vertices $v_{i-3}, v_{i-2}, v_{i-1}$ such that the vertices $\{v_{i-3}, v_{i-2}, v_{i-1}, v_i\}$ form a clique with

$$d_{v_{i-3},v_{i-2}} + d_{v_{i-2},v_{i-1}} > d_{v_{i-3},v_{i-1}}.$$

This is the topic of the next section.

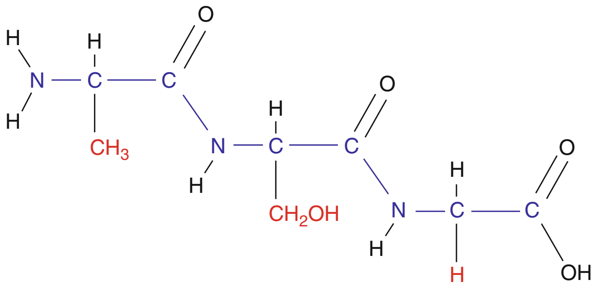
6.2 Ordering in Protein Molecules

Along with the information about the protein geometry, the NMR data provide distances between atoms as long as they are 5 angstroms (Å) or less apart. The problem becomes how to use this information to determine the coordinates of each atom of the protein molecule. The information from protein geometry tells us that the distances between atoms covalently bonded and the planar angles defined by three bonded consecutive atoms are known a priori. Clearly, the protein molecule is not a rigid structure, but these values can be considered fixed [27, 38].

This suggests a natural ordering on the atoms of the protein backbone, formed by a sequence of three atoms: N, C, C (Fig. 6.2). The protein backbone is the skeleton of the protein which already gives us a good idea of its 3D structure. For this monograph, we restrict ourselves to the protein backbone. In [14, 71], we find proposals for considering side chains (see Fig. 6.2) that distinguish between the 20 amino acids that form a protein molecule [19]. Since the distances between atoms i and $i + 3$ in the protein backbone are smaller than 5Å (in general), we can suppose that they are detected by the NMR experiments and this will provide us with the desired ordering. However, most of the NMR data are associated with pairs of hydrogen atoms [79]. An option would be to define an ordering involving just atoms of hydrogen, incorporating hydrogen atoms from the side chains, and also allowing atom repetitions in the order (Fig. 6.3).

Chemically, it does not make sense to consider two atoms in the same position, but we can do this in the ordering on the vertices of the associated graph (in fact,

Fig. 6.2 Backbone protein with side chains



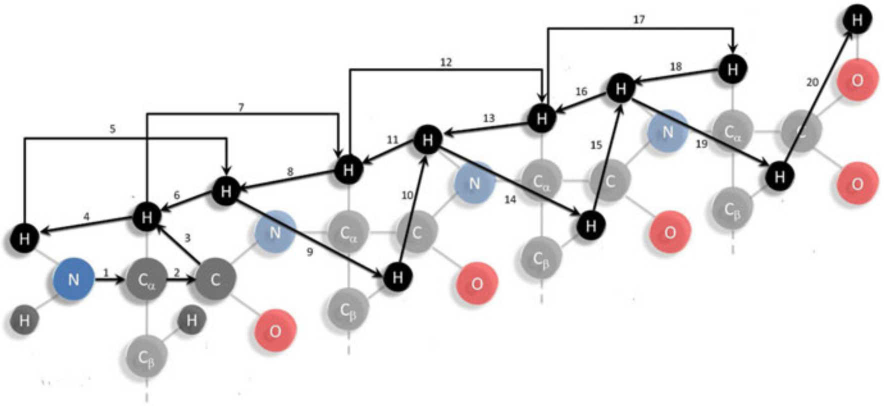


Fig. 6.3 Order on hydrogen atoms

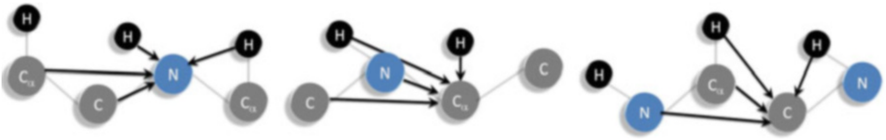


Fig. 6.4 Determination of the protein backbone using the positions of the hydrogen atoms

graph representation of a molecule is an old idea [76]). The repetition ensures that the distances $d_{i-3,i}$ are known, which may be null in some cases. From a computational viewpoint, this has an advantage because when we recalculate the position of a given repeated atom, we can verify that the numerical errors are under control [43].

Exercise 6.1 Verify in the Fig. 6.3 which pairs of atoms are repetitions.

Exercise 6.2 What happens when some of the distances, say $d_{i-1,i}$ or $d_{i-2,i}$, are null?

Suppose that, when we apply the BP algorithm, we find the positions of hydrogen atoms bonded to the protein backbone. How do we determine the positions of atoms in this chain that are of interest to us? We leave the answer to this question for the next two exercises. Remember that in Chap. 4, we saw that the intersection of four spheres, under some conditions, gives only one point.

Exercise 6.3 In the three situations depicted in Fig. 6.4, determine the quadratic systems corresponding to the intersection of four spheres.

Exercise 6.4 Show that, for each system, there exists only one solution.

There are three important aspects about the problem that we are trying to solve:

1. Distances are known (from NMR) just between close atoms,
2. Distances are known (from NMR) just between hydrogen atoms,

3. We need to solve two subproblems: (i) The calculation of the positions of the hydrogen atoms and (ii) The calculation of the positions of the atoms in the protein backbone.

   Actually, there exists another more complicated problem:

- The distances from NMR data, between neighboring hydrogen atoms, are not accurate values.

The analysis of the DMDGP considering uncertainties in the distances is a difficult problem. Some preliminaries results can be found in [4, 13, 48, 63, 74, 75]. Recall that all results that we presented in this monograph are based on the assumption that all distances are precise (real numbers), free from any error/uncertainty. However, we know that any measurements, such as those related to NMR experiments, have associated errors. In this case, we can consider that the data provided by NMR are intervals of real numbers which contain the correct distance. Even this hypothesis is an approximation of reality, since errors typically are unevenly distributed in the interval. Thus the problem is not trivial.

The good news is that this new problem provides us with an idea of how to solve Problem 3 above. We can create a new order with two main characteristics:

- We consider hydrogen atoms and protein backbone atoms at the same time,
- For the clique $\{v_{i-3}, v_{i-2}, v_{i-1}, v_i\}$, associated to the DMDGP graph, for $i = 4, \ldots, n$, all the distances $d_{i-1,i}$ and $d_{i-2,i}$ can be considered as real numbers (since they are related to bond lengths and bond angles) and just the distances $d_{i-3,i}$ are considered to have errors, modeled as intervals.

Exercise 6.5 Based on Fig. 6.5, verify that the distances $d_{i-1,i}$ and $d_{i-2,i}$ can be considered as real numbers.

Exercise 6.6 Based on Fig. 6.5, verify that some of the distances $d_{i-3,i}$ may be considered as degenerate intervals, that is, $d_{i-3,i} = 0$.

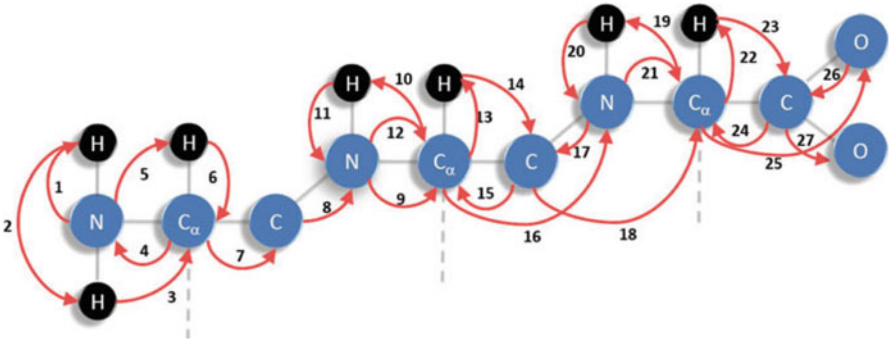


Fig. 6.5 Order with hydrogen and protein backbone atoms

Exercise 6.7 What would a BP search tree look like if we consider that the distances $d_{i-3,i}$ are intervals?

Exercise 6.8 What are the modifications necessary in the BP algorithm to incorporate interval distances?

6.3 The Polynomial Performance of the BP Algorithm

We already know that, if a given instance of the DMDGP, for all $i = 5, \ldots, n$, has an extra edge $\{v_j, v_i\} \in E$, with $j < i - 3$, such that the vertices $\{v_j, v_{i-3}, v_{i-2}, v_{i-1}\}$ generate a set of noncoplanar points, there will exist only one valid realization of the associated graph which can be computed in linear time. In general, this situation does not occur in problems related to 3D protein structures. However, we proved that under certain assumptions verified in many proteins the BP is Fixed-Parameter Tractable, which means that its exponential behavior only depend on single parameter rather than the whole size of the instance. We also verified that for several protein instances this parameter could be fixed at a constant, which suggests that the DMDGP might be a tractable problem on protein instances with exact data [56]. In part this can be explained by the fact that the protein backbone of many proteins is "tightly packed" (Fig. 6.1). The more "stretched out" the protein molecule is, the lower will be the cardinality of the pruning set E_p, causing more branches in the BP search tree.

We need to think of the BP tree as a whole and not as it is partially constructed at each step of BP, in order to have an idea of the "global behavior" of the algorithm. When the set of the pruning edges is empty, $E_p = \emptyset$, the BP tree is full, representing the entire search space. There is no difficulty in finding one solution in linear time, because it is sufficient go down the tree, by choosing any one of the two possibilities at each step of the algorithm. Since $E_p = \emptyset$, there is no possibility of errors at time of making a choice. Clearly, it is unthinkable to find all the solutions for very large n, because the solution set has cardinality 2^{n-3} (Fig. 6.6). On the other hand, suppose we have a situation described in the first paragraph of this section: for all $i = 5, \ldots, n$, there exists an extra edge $\{v_j, v_i\} \in E_p$ with $j < i - 3$. In this case, we have "only one" solution which can be found in linear time (the other one is symmetric to the plane defined by v_1, v_2, v_3), since we know what is the correct decision to be made at each step of BP (Fig. 6.7).

Increases in the computational cost of the BP algorithm are due to the required return back up the tree, when none of the calculated positions for a given vertex v is compatible with the edges $\{u, v\} \in E_p$, for $u < v - 3$ (at some previous level of the tree, a wrong decision was made). The reason that the BP algorithm is required to backtrack the tree, preventing it from an "unhindered descend" is the following:

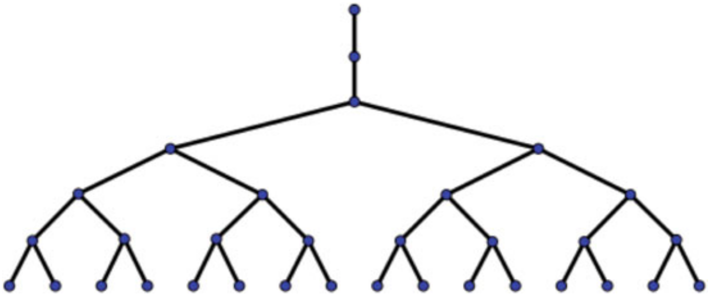


Fig. 6.6 BP tree with $E_P = \emptyset$

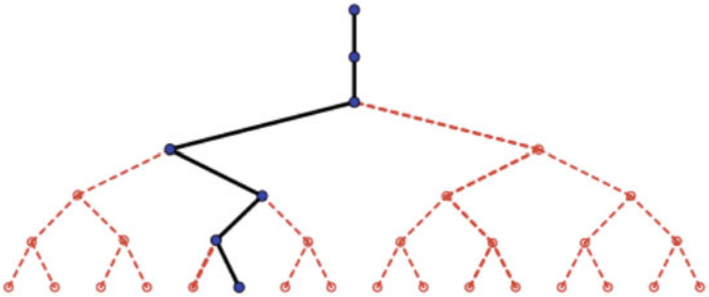


Fig. 6.7 Unique solution found in linear time

- There exists at least one vertex v_j in the DMDGP order, $v_1, \ldots, v_j, \ldots, v_n$, whose only previous vertices u, $\{u, v_j\} \in E$, are those used in the construction of the tree: $v_{j-3}, v_{j-2}, v_{j-1}$.

This means that whenever this happens, there is a duplication of the number of nodes at level j of the tree, compared to the previous level. The problem is further aggravated when there exists a set of consecutive vertices $v_j, \ldots, v_{j+k}$, for which the situation mentioned above holds, expanding the search space quickly. Suppose, for example, that at level $j = 50$ of the tree there exists $2^{20} = 1,048,576$ positions that satisfy our given data. With $k = 5$, the number of possible solutions becomes $2^{25} = 33,554,432$!

Before we make concluding remarks of this monograph, we mention that the computational cost of the BP algorithm can be reduced in at least two ways:

1. By parallelizing the algorithm [30, 64],
2. By using the concept of multiple trees [25, 69].

Chapter 7
Conclusion

We hope that this monograph inspires the reader to enter into a deeper study of Distance Geometry. Our aim was to present it as an introduction to research and teaching of the subject, where our main idea was to show that the mathematical world becomes captivating when we integrate several concepts motivated by a real and challenging problem. In exploring the area of Distance Geometry, we touched upon several different mathematical and computational fields: graph theory, geometry, algebra, combinatorics, data structures, and complexity of algorithms. We also touched upon ideas such as dimension, metric, symmetry, numerical approximation, solvability of problems and computational cost.

A fundamental topic, mainly in the applications, that was lightly considered in Chap. 6 is related to uncertainty. Chapter 6 was a different chapter from the others, because it was focused to molecular geometry which is an important field of study within Distance Geometry. We employed, implicitly and explicitly, the concepts and results from the previous chapters and also challenged the reader to solve complex problems perhaps requiring an understanding of the "subtleties" of the existent relations between mathematics and its applications.

The reader undoubtedly realizes that Distance Geometry is very rich and involves theoretical and computational challenges, as the application we selected illustrates.

C. Lavor et al., *An Introduction to Distance Geometry applied to Molecular Geometry*, SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-57183-6_7

References

1. Abud, G., Alencar, J.: Counting the number of solutions of the discretizable molecular distance geometry problem. In: Andrioni, A., Lavor, C., Liberti, L., Mucherino, A., Maculan, N., Rodriguez, R. (eds.) Proceedings of the Workshop on Distance Geometry and Applications, pp. 29–32. Universidade Federal do Amazonas, Manaus (2013)
2. Alencar, J., Lavor, C., Bonates, T.: A combinatorial approach to multidimensional scaling. In: Proceedings of the 3rd International Congress on Big Data, pp. 562–569. IEEE Computer Society (2014)
3. Alves, R., Lavor, C.: Clifford algebra applied to Grover's algorithm. Adv. Appl. Clifford Algebr. **20**, 477–488 (2010)
4. Alves, R., Lavor, C.: Geometric algebra to model uncertainties in the discretizable molecular distance geometry problem. Adv. Appl. Clifford Algebr. doi:10.1007/s00006-016-0653-2
5. Alves, R., Cassioli, A., Mucherino, A., Lavor, C., Liberti, L.: Adaptive branching in iBP with Clifford algebra. In: Andrioni, A., Lavor, C., Liberti, L., Mucherino, A., Maculan, N., Rodriguez, R. (eds.) Proceedings of the Workshop on Distance Geometry and Applications, pp. 65–69. Universidade Federal do Amazonas, Manaus (2013)
6. Andrioni, A., Lavor, C., Liberti, L., Mucherino, A., Maculan, N., Rodriguez, R. (eds.): Proceedings of the Workshop on Distance Geometry and Applications. Universidade Federal do Amazonas, Manaus (2013)
7. Bajaj, C.: The algebraic degree of geometric optimization problems. Discret. Comput. Geom. **3**, 177–191 (1988)
8. Benedetti, R., Risler, J.-J.: Real Algebraic and Semi-Algebraic Sets. Hermann, Paris (1990)
9. Billinge, S., Duxbury, P., Gonçalves, D., Lavor, C., Mucherino, A.: Assigned and unassigned distance geometry: applications to biological molecules and nanostructures. 4OR **14**, 337–376 (2016)
10. Blumenthal, L.: Theory and Applications of Distance Geometry. Oxford University Press, Oxford (1953)
11. Brünger, A., Nilges, M.: Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy. Q. Rev. Biophys. **26**, 49–125 (1993)
12. Cassioli, A., Gunluk, O., Lavor, C., Liberti, L.: Discretization vertex orders in distance geometry. Discret. Appl. Math. **197**, 27–41 (2015).
13. Cassioli, A., Bordeaux, B., Bouvier, G., Mucherino, A., Alves, R., Liberti, L., Nilges, M., Lavor, C., Malliavin, T.: An algorithm to enumerate all possible protein conformations verifying a set of distance constraints. BMC Bioinform. **16**, 16–23 (2015)
14. Costa, V., Mucherino, A., Lavor, C., Cassioli, A., Carvalho, L., Maculan, N.: Discretization orders for protein side chains. J. Glob. Optim. **60**, 333–349 (2014)

C. Lavor et al., *An Introduction to Distance Geometry applied to Molecular Geometry*, SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-57183-6

15. Crippen, G., Havel, T.: Distance Geometry and Molecular Conformation. Wiley, New York (1988)
16. Deza, M., Deza, E.: Encyclopedia of Distances. Springer, Berlin (2009)
17. Dill, K., MacCallum, J.: The protein-folding problem, 50 years on. Science **338**, 1042–1046 (2012)
18. Dokmanic, I., Parhizkar, R., Ranieri, J., Vetterli, M.: Euclidean distance matrices: essential theory, algorithms, and applications. IEEE Signal Process. Mag. **32**, 12–30 (2015)
19. Donald, B.: Algorithms in Structural Molecular Biology. MIT Press, Boston (2011)
20. Dong, Q., Wu, Z.: A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances. J. Glob. Optim. **22**, 365–375 (2002)
21. Duxbury, P., Granlund, L., Gujarathi, S., Juhas, P., Billinge, S.: The unassigned distance geometry problem. Discret. Appl. Math. **204**, 117–132 (2016)
22. Dzemyda, G., Kurasova, O., Zilinskas, J.: Multidimensional Data Visualiation: Methods and Applications. Springer, New York (2013)
23. Emiris, I., Mourrain, B.: Computer algebra methods for studying and computing molecular conformations. Algorithmica **25**, 372–402 (1999)
24. Eren, T., Goldenberg, D., Whiteley, W., Yang, Y., Morse, A., Anderson, B., Belhumeur, P.: Rigidity, computation, and randomization in network localization. Proceedings of the IEEE Infocom, pp. 2673–2684 (2004)
25. Fidalgo, F., Rodriguez, J.: Quaternions as a tool for merging multiple realization trees. In: Andrioni, A., Lavor, C., Liberti, L., Mucherino, A., Maculan, N., Rodriguez, R. (eds.) Proceedings of the Workshop on Distance Geometry and Applications, pp. 119–124. Universidade Federal do Amazonas, Manaus (2013)
26. Floudas, C., Gounaris, C.: A review of recent advances in global optimization. J. Glob. Optim. **45**, 3–38 (2009)
27. Gibson, K., Scheraga, H.: Energy minimization of rigid-geometry polypeptides with exactly closed disulfide loops. J. Comput. Chem. **18**, 403–415 (1997)
28. Gonçalves, D., Mucherino, A.: Discretization orders and efficient computation of Cartesian coordinates for distance geometry. Optim. Lett. **8**, 2111–2125 (2014)
29. Gonçalves, D., Mucherino, A., Lavor, C., Liberti, L.: Recent advances on the interval distance geometry problem. J. Glob. Optim. doi:10.1007/s10898-016-0493-6
30. Gramacho, W., Mucherino, A., Lavor, C., Maculan, N.: A parallel BP algorithm for the discretizable distance geometry problem. In: IEEE Proceedings of the Workshop on Parallel Computing and Optimization, Shanghai, pp. 1756–1762 (2012)
31. Gramacho, W., Gonçalves, D., Mucherino, A., Maculan, N.: A new algorithm to finding discretizable orderings for distance geometry. In: Andrioni, A., Lavor, C., Liberti, L., Mucherino, A., Maculan, N. Rodriguez, R. (eds.) Proceedings of the Workshop on Distance Geometry and Applications. Universidade Federal do Amazonas, Manaus, pp. 149–152 (2013)
32. Graver, J., Servatius, B., Servatius, H.: Combinatorial Rigidity. AMS, Providence (1993)
33. Grover, L.: Quantum mechanics helps in searching for a needle in a haystack. Phys. Rev. Lett. **79**, 325–328 (1997)
34. Harary, F.: Graph Theory. Addison-Wesley, Reading (1994)
35. Havel, T., Wüthrich, K.: A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of ^{1}H-^{1}H proximities in solution. Bull. Math. Biol. **46**, 673–698 (1984)
36. Havel, T., Wüthrich, K.: An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformation in solution. J. Mol. Biol. **182**, 281–294 (1985)
37. Hunt, K.: Structural kinematics of in-parallel-actuated-robot-arms. J. Mech. Transm. Autom. Des. **105**, 705–712 (1983)
38. Jackson, B., Jordán, T.: On the rigidity of molecular graphs. Combinatorica **28**, 645–658 (2008)
39. Juhás, P., Cherba, D., Duxbury, P., Punch, W., Billinge, S.: Ab initio determination of solid-state nanostructure. Nature **440**, 655–658 (2006)

40. Lavor, C.: On generating instances for the molecular distance geometry problem. In: Liberti, L., Maculan, N. (eds.) Global Optimization: From Theory to Implementation, pp. 405–414. Springer, Berlin (2006)
41. Lavor, C., Liberti, L., Maculan, N.: Grover's algorithm applied to the molecular distance geometry problem. In: Proceedings of the 7th Brazilian Congress of Neural Networks, Natal (2005)
42. Lavor, C., Carvalho, L., Portugal, R., Moura, C.: Complexity of Grovers algorithm: an algebraic approach. Int. J. Appl. Math. **20**, 801–814 (2007)
43. Lavor, C., Mucherino, A., Liberti, L., Maculan, N.: On the computation of protein backbones by using artificial backbones of hydrogens. J. Glob. Optim. **50**, 329–344 (2011)
44. Lavor, C., Liberti, L., Maculan, N.: A note on "A Branch-and-Prune Algorithm for the Molecular Distance Geometry Problem". Int. Trans. Oper. Res. **18**, 751–752 (2011)
45. Lavor, C., Lee, J., Lee-St. John, A., Liberti, L., Mucherino, A., Sviridenko, M.: Discretization orders for distance geometry problems. Optim. Lett. **6**, 783–796 (2012)
46. Lavor, C., Liberti, L., Maculan, N., Mucherino, A.: Recent advances on the discretizable molecular distance geometry problem. Eur. J. Oper. Res. **219**, 698–706 (2012)
47. Lavor, C., Liberti, L., Maculan, N., Mucherino, A.: The discretizable molecular distance geometry problem. Comput. Optim. Appl. **52**, 115–146 (2012)
48. Lavor, C., Liberti, L., Mucherino, A.: The interval branch-and-prune algorithm for the discretizable molecular distance geometry problem with inexact distances. J. Glob. Optim. **56**, 855–871 (2013)
49. Lavor, C., Alves, R., Figueiredo, W., Petraglia, A., Maculan, N.: Clifford algebra and the discretizable molecular distance geometry problem. Adv. Appl. Clifford Algebr. **25**, 925–942 (2015)
50. Lee, J., Verleysen, M.: Nonlinear Dimensionality Reduction. Springer, Berlin (2010)
51. Liberti, L., Lavor, C.: On a relationship between graph realizability and distance matrix completion. In: Migdalas, A., Sifaleras, A., Georgiadis, C., Papathanaiou, J., Stiakakis, E. (eds.) Optimization Theory, Decision Making, and Operational Research Applications, pp. 39–48. Springer, Berlin (2013)
52. Liberti, L., Lavor, C.: Six mathematical gems from the history of distance geometry. Int. Trans. Oper. Res. **23**, 897–920 (2016)
53. Liberti, L., Lavor, C., Maculan, N.: A branch-and-prune algorithm for the molecular distance geometry problem. Int. Trans. Oper. Res. **15**, 1–17 (2008)
54. Liberti, L., Lavor, C., Mucherino, A., Maculan, N.: Molecular distance geometry methods: from continuous to discrete. Int. Trans. Oper. Res. **18**, 33–51 (2010)
55. Liberti, L., Lavor, C., Alencar, J., Resende, G.: Counting the number of solutions of ^{K}DMDGP instances. Lecture Notes Comput. Sci. **8085**, 224–230 (2013)
56. Liberti, L., Lavor, C., Mucherino, A.: The discretizable molecular distance geometry problem seems easier on proteins. In: Mucherino, A., Lavor, C., Liberti, L., Maculan, N. (eds.) Distance Geometry: Theory, Methods, and Applications, pp. 47–60. Springer, New York (2013)
57. Liberti, L., Lavor, C., Maculan, N., Mucherino, A.: Euclidean distance geometry and applications. SIAM Rev. **56**, 3–69 (2014)
58. Liberti, L., Masson, B., Lee, J., Lavor, C., Mucherino, A.: On the number of realizations of certain Henneberg graphs arising in protein conformation. Discret. Appl. Math. **165**, 213–232 (2014)
59. Lima, R., Martínez, J.: Solving molecular distance geometry problems using a continuous optimization approach. In: Mucherino, A., Lavor, C., Liberti, L., Maculan, N. (eds.) Distance Geometry: Theory, Methods, and Applications, pp. 213–224. Springer, New York (2013)
60. Lindegren, L., Lammers, U., Hobbs, D., O'Mullane, W., Bastian, U., Hernández, J.: The astrometric core solution for the Gaia mission: overview of models, algorithms, and software implementation. Astron. Astrophys. **538**, 1–47 (2012)
61. Mayer-Schönberger, V., Cukier, K.: Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt, New York (2013)

62. Menger, K.: Untersuchungen uber allgemeine Metrik. Mathematische Annalen **100**, 75–163 (1928)
63. Mucherino, A.: On the identification of discretization orders for distance geometry with intervals. Lect. Notes Comput. Sci **8085**, 231–238 (2013)
64. Mucherino, A., Lavor, C., Liberti, L., Talbi, E.-G.: A parallel version of the branch & prune algorithm for the molecular distance geometry problem. In: ACS/IEEE Proceedings of the International Conference on Computer Systems and Applications, Hammamet, pp. 1–6 (2010)
65. Mucherino, A., Lavor, C., Liberti, L.: The discretizable distance geometry problem. Optim. Lett. **6**, 1671–1686 (2012)
66. Mucherino, A., Lavor, C., Liberti, L.: Exploiting symmetry properties of the discretizable molecular distance geometry problem. J. Bioinform. Comput. Biol. **10**, 1242009(1–15) (2012)
67. Mucherino, A., Lavor, C., Liberti, L., Maculan, N. (eds.): Distance Geometry: Theory, Methods, and Applications. Springer, New York (2013)
68. Nielsen, J., Roth, B.: On the kinematic analysis of robotic mechanisms. Int. J. Robot. Res. **18**, 1147–1160 (1999)
69. Nucci, P., Nogueira, L., Lavor, C.: Solving the discretizable molecular distance geometry problem by multiple realization trees. In: Mucherino, A., Lavor, C., Liberti, L., Maculan, N. (eds.) Distance Geometry: Theory, Methods, and Applications, pp. 161–176. Springer, New York (2013)
70. Porta, J., Ros, L., Thomas, F., Torras, C.: A branch-and-prune solver for distance constraints. IEEE Trans. Robot. **21**, 176–187 (2005)
71. Sallaume, S., Martins, S., Ochi, L., Gramacho, W., Lavor, C., Liberti, L.: A discrete search algorithm for finding the structure of protein backbones and side chains. Int. J. Bioinform. Res. Appl. **9**, 261–270 (2013)
72. Saxe, J.: Embeddability of weighted graphs in k-space is strongly NP-Hard. In: Proceedings of 17th Allerton Conference in Communications, Control and Computing, pp. 480–489 (1979)
73. Schlick, T.: Molecular Modelling and Simulation: An Interdisciplinary Guide. Springer, New York (2002)
74. Souza, M., Xavier, A., Lavor, C., Maculan, N.: Hyperbolic smoothing and penalty techniques applied to molecular structure determination. Oper. Res. Lett. **39**, 461–465 (2011)
75. Souza, M., Lavor, C., Muritiba, A., Maculan, N.: Solving the molecular distance geometry problem with inaccurate distance data. BMC Bioinform. **14**, S71–S76 (2013)
76. Sylvester, J.: Chemistry and algebra. Nature **17**, 284–284 (1877)
77. Thompson, H.: Calculation of Cartesian coordinates and their derivatives from internal molecular coordinates. J. Chem. Phys. **47**, 3407–3410 (1967)
78. Watson, J., Crick, F.: Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature **171**, 737–738 (1953)
79. Wüthrich, K.: Protein structure determination in solution by nuclear magnetic resonance spectroscopy. Science **243**, 45–50 (1989)
80. Wüthrich, K.: The way to NMR structures of proteins. Nat. Struct. Biol. **8**, 923–925 (2001)
81. Yemini, Y.: The positioning problem—a draft of an intermediate summary. In: Proceedings of the Conference on Distributed Sensor Networks, pp. 137–145 (1978)