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:

- $d_{1,2}, \ldots, d_{n-1,n}$: distances associated with the consecutive vertices,
- $\theta_{1,3}, \ldots, \theta_{n-2,n}$: planar angles associated with three consecutive vertices,
- $\omega_{1,4}, \ldots, \omega_{n-3,n}$: 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 $d_{1,2}, \ldots, d_{n-1,n}$ are, obviously, obtained from the definition of the DMDGP, and the values $\theta_{1,3}, \ldots, \theta_{n-2,n}$ 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 $(i = 4, \ldots, n)$ [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?

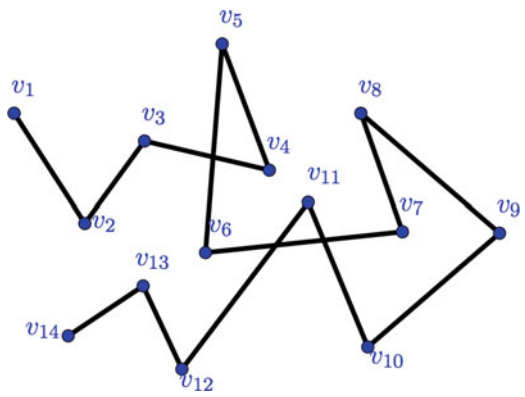

Fig. 5.1 DMDGP instance as a molecule

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:

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

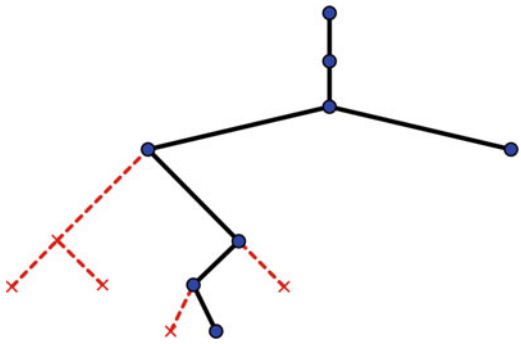

Fig. 5.2 Solution obtained by BP algorithm

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

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

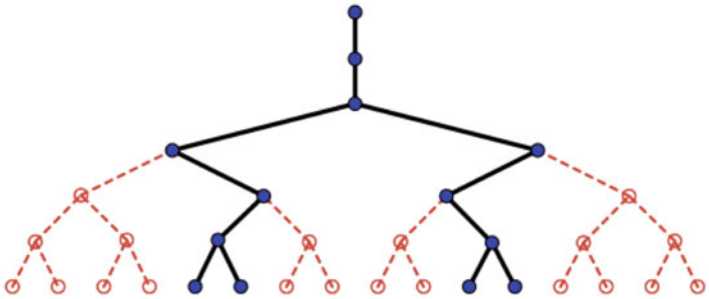

Fig. 5.3 Symmetric solutions

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?