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

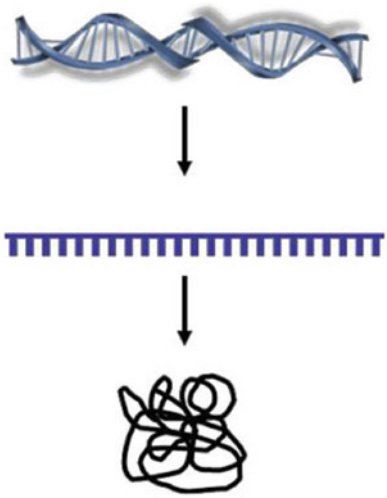

Fig. 6.1 DNA and protein

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,

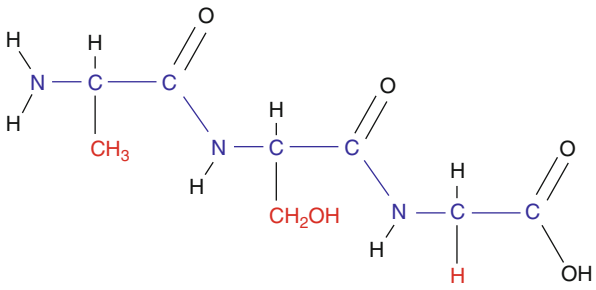

Fig. 6.2 Backbone protein with side chains

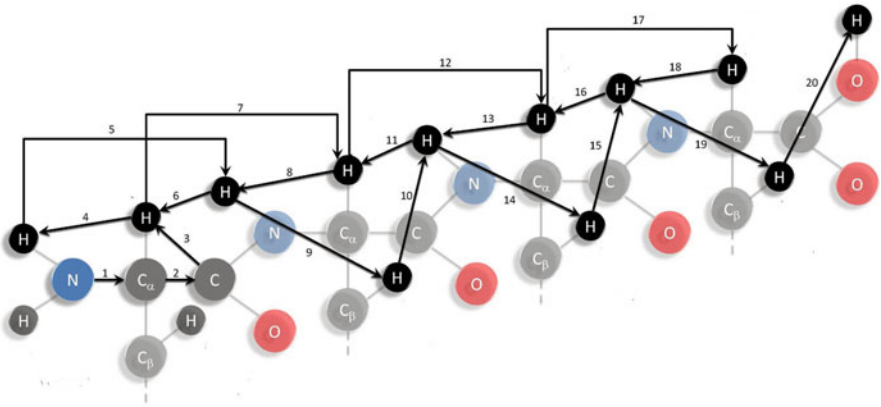

Fig. 6.3 Order on hydrogen atoms

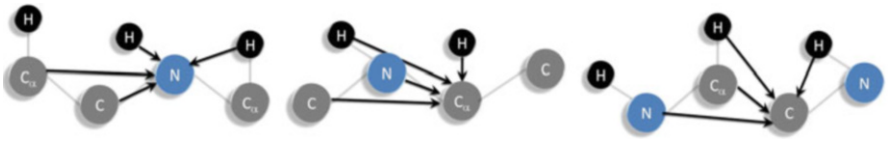

Fig. 6.4 Determination of the protein backbone using the positions of the hydrogen atoms

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, $d_{i-3,i} = 0$.

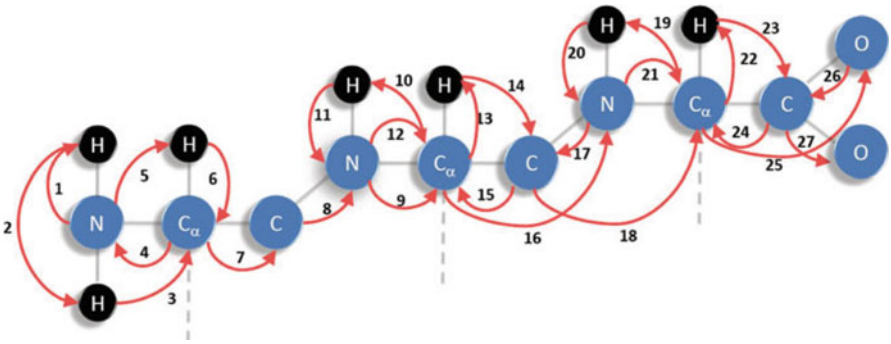

Fig. 6.5 Order with hydrogen and protein backbone atoms

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:

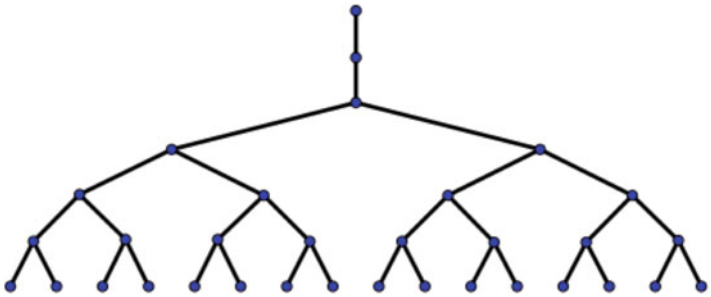

Fig. 6.6 BP tree with $E_P = \emptyset$

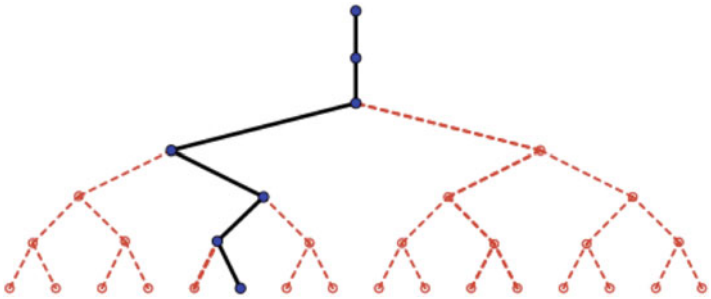

Fig. 6.7 Unique solution found in linear time

- 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].