Chapter 16
Distance Geometry in Structural Biology: New Perspectives

Thérèse E. Malliavin, Antonio Mucherino, and Michael Nilges

Abstract Proteins are polypeptides of amino acids involved in most of the biological processes. In the last 50 years, the study of their structures at the molecular level revolutioned the vision of biology. The three-dimensional structure of these molecules helps in the identification of their biological function. In this chapter, we focus our attention on methods for structure determination based on distance information obtained by nuclear magnetic resonance (NMR) experiments. We give a few details about this experimental technique and we discuss the quality and the reliability of the information it is able to provide. The problem of finding protein structures from NMR information is known in the literature as the molecular distance geometry problem (MDGP). We review some of the historical and most used methods for solving MDGPs with NMR data. Finally, we give a brief overview of a new promising approach to the MDGP, which is based on a discrete formulation of the problem, and we discuss the perspectives this method could open in structural biology.

16.1 Introduction

The vision of biology has been fundamentally modified during the second part of the twentieth century by the analysis of the cell function at the molecular

T.E. Malliavin (✉)
Institut Pasteur, Paris, France
e-mail: therese.malliavin@pasteur.fr

A. Mucherino
IRISA, University of Rennes 1, avenue de General Leclerc, Rennes 35042, France
e-mail: antonio.mucherino@irisa.fr

M. Nilges
Institut Pasteur, Paris, France
e-mail: michael.nilges@pasteur.fr

A. Mucherino et al. (eds.), *Distance Geometry: Theory, Methods, and Applications*,
DOI 10.1007/978-1-4614-5128-0_16,

level. This brought about a molecular description of the interactions between the molecular agents (biomolecules) which perform important biological processes. Just to mention some examples, molecular motors are the essential agents of movement in living organisms, transcription factors regulate the genetic expressions, enzymes are able to catalyze chemical reactions, and ion channels help establishing and controlling the voltage gradient across the cell membrane. Moreover, transport proteins perform the function of moving other materials inside an organism.

The description of biomolecules at molecular level has been possible for 50 years, due to the development of methods to study the molecular structure of biomolecules. Indeed, these structures are essential in order to understand the function they are able to perform. The slightest modifications in this structure can drastically change the corresponding biomolecular function, as it is encountered, for example, for neurodegenerative diseases [21].

Biomolecular structures can be studied at different levels. A protein is a polypeptide of *amino acids*, named *protein residues* when they are inserted into the polypeptide chain. Polypeptide synthesis is performed through the controlled formation of a peptide bond between two amino acids, where each amino acid pair loses a water molecule. The protein main chain is usually referred to as *backbone*, whereas the atoms that are specific for each residue form the so-called *side chains*.

Proteins display a hierarchical level of organization. Their *primary structures* consist of the sequence of amino acids composing the molecule. Amino acids bond to each other to form a chain, which, under chemical and physical forces, gives rise to three-dimensional structures that are specific for a given primary structure. The *secondary structures* of proteins represent local arrangements of residues: in α-helices, the backbone is arranged as a helix, whereas in β-sheets, the structure is formed by strands of residues over a common plane. The *tertiary structure* is the global arrangement of the amino acids of monomeric proteins, for which there is a unique sequence of amino acids. For more complex proteins, the global three-dimensional structure is given by their *quaternary structure*, build up from the tertiary structures of the various chains of amino acids which can compose the molecule.

In the following, the organization in the three-dimensional space of the atoms of a molecule will be referred to as *conformation* of the molecule. This conformation, together with its chemical architecture, will be referred to as *structure* of the molecule. In the literature, papers may refer to conformations or structures, but in the problem presented here, the actual unknown is the conformation, because the protein structure can be deduced from its conformation, plus its chemical composition.

The focus of this chapter is on methods and algorithms for the identification of the three-dimensional conformations of proteins. The rest of this chapter is organized as follows. Our discussion begins in Sect. 16.2 with the different possible representations for protein conformations that can be considered when solving problems related to such molecules. This representation is strongly related to the complexity of considered solution methods. In Sect. 16.3, we introduce structural biology and discuss its importance for understanding biological processes. Then, we will focus our attention on nuclear magnetic resonance (NMR) experiments and on methods for finding protein conformations from NMR data. In Sect. 16.4, we will give an overview of NMR experiments, and we will discuss about possible sources of errors that may affect the distance information that it is able to provide.

In Sect. 16.5, we will introduce the molecular distance geometry problem (MDGP) and we will study its complexity under different hypotheses. In Sect. 16.6, we will present some basic techniques for refining the distance information given by NMR. In Sect. 16.7, we will briefly present the first method that was used for solving MDGPs with NMR data, and we will mention to some of the issues that caused its replacement with global optimization techniques. A discussion on global optimization for the MDGP is given in Sect. 16.8. The most currently used technique for solving MDGPs by optimization is based on the meta-heuristic simulated annealing (SA): most protein structures that are currently available on the protein data bank (PDB) and that have been analyzed by NMR experiments were obtained by some SA-based global optimization computational tools. We will briefly present the basic idea behind this approach, as well as a new deterministic approach that is based on a discretization of the problem. Finally, in Sect. 16.9, we will give some directions for future research.

16.2 Protein Representation

We begin our discussion on methods and algorithms for protein structure determination from NMR data with a short overview of suitable representations for protein conformations. We refer to [40], where a similar discussion was presented in a different context.

An atom can be represented by the coordinates of its mass center in three-dimensional space. Therefore, if a molecule is simply seen as a set of atoms, a possible representation is given by a set of coordinates in the space. This representation is known as *full-atom* representation of a molecule, which involves $3n$ real variables, three for each of the n atoms forming the molecule. Other more efficient representations can be however employed for molecular conformations.

The task of finding an efficient representation is evidently easier when information is available on the chemical composition of the molecule. As already remarked, proteins are chains of amino acids, and the subgroup of atoms that is common to each residue forms the so-called protein backbone. Among the atoms contained in this backbone, more importance is given to the carbon atom usually labeled with the symbol C_α. In some works in the literature (see for example [24, 37]), this C_α atom is used for representing an entire residue. In this case, therefore, the protein conformation is represented through the spatial coordinates of n_{aa} atoms, where n_{aa} is the number of residues forming the protein. Considering that each residue can contain 10–20 atoms, it is clear how simplified this representation is. The sequence of C_α atoms is also called *trace* of the protein. We remark that this representation cannot be employed for discriminating among the 20 different amino acids that can make up the protein.

More accurate representations of protein backbones can be obtained if more atoms are considered. If, together with the carbon C_α, two other atoms, which are bonded to this C_α, are also considered (another carbon C' and a nitrogen N), then the whole protein backbone can be reconstructed. In other words, the coordinates of three atoms per amino acid are sufficient for representing a whole protein conformation without side chains. Therefore, a protein backbone can be represented precisely by a sequence of $3n_{aa}$ atomic coordinates.

This representation is however not much used, because there is another representation for the protein backbones which is much more efficient. Four consecutive atoms in the sequence of atoms N, C_α, and C representing a protein backbone form a *torsion angle*, i.e., the angle between the plane formed by the first triplet of atoms and the plane formed by the second triplet of atoms in this quadruplet. Torsion angles can be computed from available atomic coordinates, and, since some distances are angles between bonded atoms are known, the procedure can also be inverted. The representation of a protein which is based on torsion angles is more efficient, because the protein backbone is described by fewer variables.

In the applications, the representation based on torsion angles is further simplified. The sequence of atoms on the protein backbone is a continuous repetition of the atoms N, C_α, and C. Each quadruplet defining a torsion angle contains two atoms of the same kind that belong to two bonded amino acids. Then, the torsion angles can be divided into three groups, depending on the kind of atom that appears twice. Torsion angles of the same group are usually denoted by the same symbol: the most used symbols are ϕ, ψ, and ω. The torsion angle ω is rather constant and its value is generally very close to 180°: because of the resonance stabilization of the amide (peptide) bond and because of the carbonyl double bond, the four involved atoms, C_α, C' (belonging to the first amino acid), and N, C_α (belonging to the second one in the sequence), are constrained to be on the same plane. Therefore, a protein backbone can be represented by a set of $2n_{aa} - 2$ variables, one for each torsion angle ϕ and ψ that can be defined. There is a variant of the SA-based algorithm described in Sect. 16.8.1 that is based on this protein representation.

When the whole protein conformation needs to be represented, other torsion angles (usually denoted χ) are defined for the description of the amino acid side chains, where each of such subsequences of angle χ is specific to the different chemical properties of the amino acids.

Section 16.8.2 is devoted to a novel approach to distance geometry where the problem is discretized. In this case, the variables employed for the protein representation do not need to vary in a continuous space, but they can take a finite number of values. In the simplified case in which there is no uncertainty in the input data [42], a protein can be represented by a vector of binary variables. Otherwise, a vector of integer values can be used for the representation. Naturally, these discrete representations rely on a priori known information on proteins (as in the case of the torsion angle representation). Additional details about these discrete representations are given in Sect. 16.8.2.

For evident reasons, there is no best representation. A representation needs to be selected on the basis of the properties of the problem to be studied. In distance geometry for structural biology, the torsion angle representation is the most currently used one. In different situations, however, other representations could be more appropriate.

16.3 Importance of Structural Biology to Understand Biological Processes

The conformation and the structure of biomolecules are very important to understand their function and to analyze possible interactions of the molecule with other molecules, helping in this way the development of new drugs [52]. Because of the essential role of the molecular structure of molecules, a scientific field, called *structural biology*, whose main aim is to identify and study biological structures, has experienced an enormous development.

Structural biology originated from the application of powerful physical techniques to biological objects. It has also largely benefited from the development of molecular biology biochemistry and cellular biology techniques. The widespread application of structural biology methods has produced a quite astonishing molecular description of life. Due to this great impact, structures of biological molecules are deposited in public web databases. The most important database is named PROTEIN DATA BANK (PDB: http://www.rcsb.org) [2], which reached the number of 80,000 deposited molecules at the beginning of 2012.

Biomolecular structures can be investigated at several levels: single biomolecules, biomolecular complexes and assemblies, and cellular organs. Different methods can be applied for studying these biomolecular structures. In general, the information provided by such methods concerns the molecular electronic density and the spatial proximity information. The molecular electronic density describes the position of electronic clouds in molecules. Information on spatial proximity corresponds to the measurement of distances (or angles) between atoms or regions of the molecule.

NMR is one of the major techniques used for studying biomolecular structures. NMR is able to give very sensitive information on distances or angles between atoms. NMR is also able to provide information on the internal dynamics of the molecule. There are also other experimental techniques that can give measurements concerning distances between atoms: an example is given by fluorescence techniques (FRET, cellular imaging) and by hybrid methods, such as mass spectrometry coupled with cross-linking. Due to the very high sensitivity of fluorescence and to the lack of limitation on the size of the analyzed objects, these techniques are likely to continue their development in coming years.

Historically, distance-based methods started to develop when the methods based on electronic density were already well-established. The main objective of such methods was to identify a three-dimensional conformation for a molecule from the experimental data obtained while applying the mentioned experimental methods (for a detailed definition of this problem, see Sect. 16.5). The development of these distance-based methods represented a new challenge in biology and required an intensive intellectual investment.

Other problems in structural biology can also benefit from distance information. One important example is the docking problem, where the conformation of two (or more) molecules, during their interaction, is searched. In general, it is supposed that the conformation of the first molecule M_1, as well as the conformation of the second molecule M_2, is known. The interest is in discovering the way M_1 and M_2 arrange their conformations in space during the interaction. NMR and other techniques, such as FRET, can provide distance information between pairs of atoms (u,v) such that u belongs to M_1 and v belongs to M_2. Since the conformations of M_1 and M_2, separately, are supposed to be known, other distances can be derived, and, together with the measured distances, can be exploited for analyzing the interaction between the two molecules [22]. In docking, generally, M_1 represents a protein, whereas M_2 is generally referred to as the ligand and can be another type of molecule.

Homology modeling or, more accurately, comparative modeling [49] attempts the construction of protein conformations by using their chemical composition, which can be easily derived from the sequence (or the sequences) of amino acids forming the molecule, and the similarity of the sequence with other proteins of known conformation. The geometric information required for the structure construction is obtained by comparing the sequence of the protein under study to the sequences of proteins with known structures. The idea is to associate to the atoms of the protein under study some geometric constraints so that they can resemble the local conformation of a protein having a similar sequence. Then, a distance-based method can be used for predicting the conformation of the protein under study.

16.4 Geometric Parameters Measured by Nuclear Magnetic Resonance in Biomolecules

NMR studies the behavior of the magnetic moments of spin nuclei and is based on the observation of the so-called NMR resonance. A resonance frequency characterizes the return of each perturbed magnetic moment to equilibrium. In proteins, the nuclei ^{1}H, ^{13}C, and ^{15}N can be observed. The protein sample is submitted to an intense external magnetic field, inducing the alignment of the magnetic moment of the observed nuclei. The perturbation of any aligned magnetic moment is transmitted through dipolar interactions of the moments to the magnetic moments of neighboring nuclei. The transmission of the perturbation is called nuclear Overhauser effect (NOE) and is roughly proportional to d^{-6}, where d is the distance between two protons belonging to the two atoms u and v. The NOE between u and v is located in the spectrum at coordinates δ_u and δ_v representing the *chemical shifts* of u and v. These chemical shifts are deduced from the corresponding resonance frequencies.

Errors in the measurement of the distances by NMR can have a number of reasons:

- The sample molecule can undergo dynamics or conformational exchange so that the conversion of the measured signal into a distance becomes difficult.
- The signal recorded during NOE measurement may be distorted by experimental noise or by processing artifacts.
- The NOE measurement related to two atoms u and v is also influenced by other neighboring atoms through a process called *spin diffusion*.

The most common procedure to minimize the effects of spin diffusion and internal mobility is to qualitatively classify the NOE intensities by converting them into distance intervals [25]. A strong NOE is typically assigned to an interproton distance below 2.7 Å, a medium NOE to a distance below 3.3 Å, and a weak NOE to a distance below 5.0 Å [5]. These values define therefore the upper bounds on the possible distances associated to the pair of atoms. The corresponding lower bounds are defined by the sum of the Van der Waals radii of the involved atoms. Thus, from an NOE measurement, a suitable interval can be defined, where the actual distance d between the two atoms u and v is (most likely) contained.

It is important to remark that not all obtained interval distances can be unambiguously assigned to a pair of atoms (u,v). More than one pair of atoms can have the same chemical shifts δ_u and δ_v so that one single NOE measurement (one real value) can refer to several distances. For a set of undistinguished pairs (u,v), the relationship between NOE measurement and the distances $d(u,v)$ is approximately

$$\bar{d} = \left[\sum_{(u,v)} [d(u,v)]^{-6} \right]^{-1/6} .$$

In this case, the assignment of a distance to a pair (u,v) is generally done in an iterative way. Often, only unambiguous NOEs are used at first in order to identify a possible conformation for the protein under study. Then, additional NOEs can be assigned on the basis of some preliminary found three-dimensional conformations [17]. The ambiguous NOEs can be also automatically managed [45, 47] during the structure calculation.

Forms of ambiguity are common in methylene and propyl groups of valines and leucines. In this situation, the distance constraints are often directed to a pseudoatom [56]. A pseudoatom can be placed halfway between the two atoms of a methylene group: the distances concerning the real atoms are successively increased with respect to the ones obtained for the pseudoatom. Pseudoatoms can also be used to describe unresolved NOEs involving protons that are equivalent due to motion, such as protons in methyl groups or aromatic rings. The necessary correction for the

NOE-derived distances can be deduced from theoretical considerations [26], as well as from the size of the rotating group, for example, the upper bound for an NOE involving a methyl group is often increased by 1 Å [56].

The chemical shift strongly depends on the type of the nucleus (^{1}H, ^{13}C, or ^{15}N) and on the chemical environment. The latter observation led to an empirical relationship in order to correlate the chemical shifts of C_α and H_α atoms to the secondary structures to which the corresponding amino acid belongs. This correlation is commonly named chemical shift index (CSI) [54]. TALOS+ [51] is a software tool based on a neural network that is able to predict the secondary structures of subsequences of amino acids from chemical shifts obtained by NMR experiments. In practice, this software tool is able to provide some constraints on the ϕ and ψ torsion angles which are generally employed for the protein backbone representation (see Sect. 16.2).

NMR usually provides only short-range distances. If two atoms are more than 5–6 Å apart, then there is no NOE signal that can be measured for estimating their relative distance. Furthermore, only intervals between pairs of atoms which are visible on the NMR spectra can be estimated. Only distances between pairs of hydrogen atoms are useful for structure determination of biological macromolecules.

This makes it necessary to complement the NOE information by additional information derived from the local geometry of the molecule. If the chemical structure of the molecule is known, distances between bonded atoms or angles among triplets of bonded atoms can be computed. In general, chemical bonds allow some small variations on these relative distances, but it is very common to fix such distances for reducing the degrees of freedom of the whole molecular structure. The exact values for these distances can be obtained, for example, by X-ray crystallography measurements of small molecules or single amino acids [12]. The use of such distances makes it possible to consider the torsion angle representation of proteins discussed in Sect. 16.2. Although this is the mostly common approach, there is a quite original approach where the atoms are isolated [14]; here, NMR distances and bond distances play the same role.

16.5 The Molecular Distance Geometry Problem

The information provided by NMR essentially consists in a list of distances between some pairs of atoms of the considered molecule. NMR is also able to provide additional information, such as lower and upper bounds on the backbone torsion angles. However, this additional information can be converted in distance-based constraints, so that we can consider, in general, that the available information about the molecule only consists of distances.

The distance geometry problem (DGP) is therefore the problem of identifying the three-dimensional conformation of a molecule by exploiting a set of available distances between some pairs of its atoms [9, 18]. Formally, we can represent an instance of the DGP as a weighted undirected graph $G = (V, E, d)$ having the following properties. The vertex set $V = \{1,2,\dots,n\}$ contains vertices v representing atoms (i.e., the protons of the atoms) which compose the molecule, in a certain predefined ordering. In the following, the cardinality of V, i.e., the number of atoms/vertices in the graph, will be referred to as n or $|V|$. The edge set E contains all pairs of vertices (u,v) for which the distance between the atoms corresponding to u and v is known; the weight $d(u,v)$ associated to the edge (u,v) provides the numerical value of the distance. It can be an exact value (i.e., one single real numerical value), or, more often, an interval. Finally, we suppose that a total order relation is associated to the vertices of G, which may not correspond to the natural atomic ordering in some molecules such as proteins. When molecules are concerned, the DGP is usually referred to as molecular DGP (MDGP). With a little abuse of notation, we will refer to each $v \in V$ as "vertex" of G, as well as "atom" of a molecule.

At the beginning of this discussion, we will suppose that all distances in G are precise. In this case, the MDGP can be seen as the problem of finding a conformation $x = (x_1, x_2, \dots, x_n)$ such that all constraints

$$||x_u - x_v|| = d(u,v) \qquad \forall (u,v) \in E \tag{16.1}$$

are satisfied. In the formula, $||\cdot||$ represents the computed distance between two atomic coordinates belonging to the conformation x, whereas $d(u,v)$ represents the known distance between the two atoms (the weight associated to the edge). The MDGP is a constraint satisfaction problem.

Let us suppose that the distance between all pairs of atoms u and v is known. In such a case, the number of equations (16.1) is $n(n-1)/2$ (naturally, two edges (u,v) and (v,u) correspond to the same distance). In order to fix the conformation in the three-dimensional space (for avoiding to consider solutions that can be obtained by translating and/or rotating other solutions), the first three atoms of the molecule can be fixed in space. At this point, there are other $3n-9$ atomic coordinates to identify in order to find a conformation x which satisfies all constraints (16.1). Note that the number of coordinates is one order smaller than the number of distances. Therefore, in the simple case in which all interatomic distances are available, the distance information is redundant. In other words, only a subset of distances is actually necessary for finding a solution x to the MDGP.

Let us suppose that we need to find the position in space for the atom $v \in V$, and that all the other atoms that precede v in the ordering associated to V have already been positioned. If all distances are available, in particular the distance between $v-1$ and v is available. Geometrically, the constraint (16.1) associated to this distance defines a sphere which is centered in the position of $v-1$ and has radius $d(v-1,v)$. Hence, the possible positions for v belong to this sphere. Since similar spheres can be defined for all the other atoms u such that $u < v$, the coordinates for v can be identified by intersecting all these spheres. As it is well known, in the hypothesis that all distances (radius) are precisely known, the intersection between two spheres gives one circle, the intersection among three spheres gives two points, and the intersection among four spheres gives one point only. As a consequence, any additional sphere to be intersected with the others would not produce any additional information. It is important to remark that there exist particular cases where the sphere intersections can provide different results (e.g., the intersection of three spheres with aligned centers gives a circle and not two points), but the possibility for this to happen has probability 0 in a mathematical sense.

Let G be a graph representing an instance of the MDGP. If we suppose that the first four vertices are placed in fixed positions and that, for each other vertex $v > 4$, there are four adjacent vertices, i.e., four edges (u,v) with $u < v$, then the MDGP can be solved in linear time and there is only one possible solution. In this hypothesis, indeed, for each $v > 4$, there are at least four spheres that can be intersected for the identification of the coordinates of v. Since the intersection always produces one point only, the unique solution to this MDGP can be found in linear time. We remark that graphs G satisfying this property for the edges in E are called *trilateration graphs*, and it has been formally proved that MDGPs related to trilateration graphs can be solved in polynomial time [13]. There is, in fact, a solution method for the MDGP with exact distances that is based on this hypothesis [10, 55].

In general, however, one is far from this ideal situation. As discussed in Sect. 16.4, the quantity of distances estimated by NMR is limited to short-range distances, and they mainly concern pairs of hydrogen atoms, while a protein is composed by hydrogens but also carbons, nitrogens, oxygens , and some sulfur. Moreover, NMR distances are estimated and not measured precisely. In general, therefore, the MDGP is an NP-hard problem [50]. We will discuss in the next session some suitable techniques for refining NMR distances and for generating additional distances from the ones obtained by NMR.

16.6 Refinement of NMR Distances

As previously discussed, NMR experiments are able to provide a list of distances for a subset of atom pairs from a given molecule. The majority of such distances are imprecise, i.e., they are represented by suitable intervals where the actual distance is supposed to be contained. Moreover, the distances are generally not available for all pairs of atoms, but rather only a small subset of distances can be estimated by NMR.

Before any method for the solution of the MDGP can be applied, it is very important to verify the quality of the distances that were obtained by NMR. An effective and simple test for verifying whether the distances are compatible to one another is the one employing the well-known *triangle inequality*. Suppose there are three vertices u, v, and w such that the three edges (u,v), (v,w), and (u,w) are present in the edge set E. The three vertices form the triangle $\widehat{uvw}$, and the triangle inequality

$$d(u,w) \leq d(u,v) + d(v,w) \tag{16.2}$$

ensures that one side of the triangle ((u,w) in this case) is not larger than the sum of the other two. If the triangle inequality is not satisfied, then some of the involved distances need to be corrected. If there are distances not satisfying the triangle inequalities, then the solution to the MDGP cannot be a Euclidean object, which contradicts the definition of molecular conformation. This compatibility test can be easily generalized to interval distances: in this case, the portion of interval associated to the distance $d(u,w)$, where there are distances not satisfying the inequality (16.2), can be discarded.

Let us suppose now that we have the three vertices u, v, and w and that the edge (u,w) is not available. In this case, we can estimate the distance associated to this pair of vertices u and w by exploiting the triangle inequality. Because of Eq. (16.2), the distance $d(u,w)$ has as upper bound the sum of the two distances associated to (u,v) and (v,w). If the upper bound is considered for the distance, we have a degenerate triangle $\widehat{uvw}$ (the angle in v is equal to 180°). Smaller values for the distance produce different triangles with different values for the angles in v. However, there is another limit on the values for the distances, which is the one corresponding to an angle in v equal to 0°. Therefore, the lower bound for the distance is

$$d(u,w) \geq |d(u,v) - d(v,w)|. \tag{16.3}$$

This lower bound can be increased in case it is smaller than the sum of the Van der Waals (VdW) radii of the two involved atoms u and v.

Based on these simple rules, there is a procedure for reducing the interval lengths obtained by NMR and for redistributing the distance information along the atoms of the considered molecule. It is generally called *bound smoothing* procedure. In the very first works on this topic [3, 9], the term distance geometry described this procedure, checking the consistency of a set of distance intervals. Only later it became the preprocessing step for the following structure generation step. Bound smoothing allows to reduce the distance intervals before attempting the solution to the problem, and, at the same time, it allows to distribute the NMR information, originally mainly concerning hydrogen atoms, to other atoms of the molecule.

Since there are n^3 possible triangles in a molecule formed by n atoms, the bound smoothing procedure can be quite expensive. However, the NMR information is rather sparse, and therefore not all triangles actually have to be checked. The search for the possible triangles can be optimized by considering that, for each given edge (u,w), only pairs of edges (u,v) and (v,w), for some $v \in V$, are of interest. We remark that, in graph theory, these triangles are named *cliques* and that the enumeration of all cliques of a graph G with a predefined size K (in our case, $K = 3$) can be performed in polynomial time.

Apart from the triangle inequalities, there are other higher-order inequalities that can be verified in order to have the compatibility among the distances in G. Tetrangle, pentangle, and hexangle inequalities involve the distances between four, five, and six atoms, respectively. These inequalities can, in theory, be used just like the triangle inequalities in the bound smoothing procedure. They are actually able to better refine the NMR distances, by reducing the difference between the lower and the upper bound in the intervals which represent the distances. Whereas the triangle inequality is valid for all dimensions of space, some of these additional inequalities are more specific to three-dimensional space [9]. However, the computational cost increases for the verification of these higher-order inequalities. For this reason, only the triangle and the tetrangular inequalities have been employed in the past [11]. Nowadays, thanks to the increasing computer power, higher-order inequalities might also be considered.

16.7 The Metric Matrix Distance Geometry

The metric matrix distance geometry (MMDG) has been the first employed method for NMR structure determination [4, 7–9, 18, 19]. In the following, we will give a few details about this method and discuss the reason why it was discarded in recent years. The interested reader can refer to [1] for additional details about the MMDG. The basic idea is to exploit the properties of a matrix of interatomic distances, to which we refer as *metric matrix*, and to perform the following four main steps:

1. The available NMR distances are checked for consistency and refined by applying a *bound smoothing* procedure, as discussed in Sect. 16.6.
2. For any NMR distance which is represented by an interval, one sample distance is chosen (this process is called *metrization process* when this choice is made consistently).
3. The metric matrix is derived from the distance matrix; the eigenvectors and eigenvalues of the metric matrix are computed: this allows to generate the coordinates of the atoms forming the molecule (*embedding phase*).
4. Possible errors in the obtained conformation are corrected by applying optimization techniques (*optimization phase*).

As discussed in Sect. 16.6, bound smoothing is an important preprocessing step for the solution of an MDGP containing NMR data. However, the resulting set of distances is such that many distances are still represented by intervals (whose length can be up to 3 Å or more). This is the reason why the MMGP has an additional preprocessing (Step 2), where sample distances are chosen from the intervals.

In the simplest implementation, each distance is randomly chosen in each single interval, independently from the choices made for other intervals. However, there is an important issue regarding this simple process. After step 1, all interval distances are compatible to each other: all possible triangles (cliques) satisfy the triangle inequalities. Once three exact distances have been chosen in step 2, however, the corresponding triangle inequality will not be satisfied anymore. To overcome this difficulty, one can use metrization, a process where the bound smoothing is repeated after each distance choice.

Another important aspect of the metrization is given by the ordering in which the sample distances are chosen. If the natural order for the atoms is chosen (we consider the distances related to the first atom, and then we proceed with the ones related to the successive atoms, until the end), then there is the risk of introducing artifacts. Empirically, it was shown that the better results are obtained when the sequence of intervals is randomly chosen [27].

Once a set of exact distances is computed from the available intervals, the metric matrix $\mathscr{G}$ can be defined (see [1] for additional details). Then, the eigenvalues and the eigenvectors of $\mathscr{G}$ are computed. The eigenvectors provide the coordinates of the atoms forming the molecule, i.e., they provide the solution to the MDGP.

Unfortunately, this step of the MMDG does not always provide an acceptable result. The exact distances taken from the intervals during the metrization step, indeed, may not be consistent with a three-dimensional Euclidean object (such as a protein conformation). In this case, the number of eigenvalues which is greater than 0 is $k > 3$ so that our conformation does not belong to the three-dimensional space.

The exact distances taken from the intervals during the metrization step very likely are not consistent with a three-dimensional Euclidean object (such as a protein conformation). In this case, the number of nonzero eigenvalues is $k > 3$ so that our conformation does not belong to the three-dimensional space. One usually truncates therefore the eigenvalue series after the third. This is the optimal projection of the higher-dimensional object into three-dimensional space. In some cases eigenvectors related to more than three strictly positive eigenvalues need to be considered [53].

The last step of the MMGP method consists in *optimizing* the conformation obtained in the embedding step. The generic approach is to define a penalty function which gives penalties to violations for the available constraints, as well as for some local conformations that are not typical in proteins. The chosen penalty function can be minimized, for example, by using a conjugate gradient minimization method [44], which is a local search optimization method.

The MMGP was initially employed in structural biology for solving MDGPs with NMR data. It has largely been replaced by these methods, because of convergence issues and since optimization algorithms are more flexible. More recent approaches to the MDGP are based on suitable reformulations of the MDGP as a global optimization problem.

16.8 Methods Based on Global Optimization

Global optimization aims at finding the global minimum (or the set of global minima) of a certain mathematical function (called *objective function*) under the hypothesis that some constraints are satisfied. Many real-life applications lead to the formulation of a global optimization problem [38]. Depending on the properties of the objective function and constraints, suitable methods can be employed for the solution of the optimization problem.

The MDGP can be reformulated as an unconstrained global optimization problem. The satisfaction of the constraints based on the distances can be measured by computing the difference between the left and the right side in the constraints (16.1). In order to verify the overall satisfaction of the available constraints, a penalty function can be defined, whose general term is related to the generic constraint. Different penalty functions can be defined for the MDGP, and the most used one is the largest distance error (LDE):

$$\mathrm{LDE}(x) = \frac{1}{|E|} \sum_{(u,v)} \frac{|\,||x_u - x_v|| - d(u,v)\,|}{d(u,v)}. \tag{16.4}$$

Finding the global minimum of this penalty function allows to obtain solutions to the MDGP. If all distances are compatible to each other and there are no errors, the LDE value in conformations x which are solution for the MDGP is supposed to be zero.

Penalty functions that can be defined for the MDGP generally contain several local minima. This makes the task of finding the global minimum (or the global minima) of the penalty function very difficult. Many methods may get trapped at local minima, and there might not be ways to verify whether the found minimum is global or not. There is a wide literature on global optimization, the interested reader is referred, for example, to [20].

In the following two sections, we discuss some methods for the MDGP which are based on a global optimization reformulation of the problem. We point out that this is not meant to be a comprehensive survey. We rather focus the rest of this chapter on two particular methods: the one which is nowadays mostly used for the determination of the protein conformations deposited on the PDB (see Sect. 16.8.1) and another one that is more recent and potentially able to identify better-quality conformations of proteins (see Sect. 16.8.2). The reader who is interested in a wider discussion on global optimization methods for the MDGP is referred to recent surveys [30, 34, 36].

16.8.1 SA-Based Methods

The SA [23] was introduced in 1983 by Kirkpatrick in order to solve nonlinear optimization problems. The basic idea is to simulate the annealing physical process, where a given system (such as a glass of water) is cooled down slowly for obtaining a low-energy structure (such as the crystalline structure of a piece of ice). In the simulation, particles of the physical system are represented by the variables of a certain objective function, while its energy is given by the objective function value. Randomly generated solutions to the problem are computed during the simulation, and, as the system is cooled down, the possibility to accept solutions that increase the system energy gets lower and lower. SA depends on a set of parameters, such as the initial temperature, the cooling schedule, and the number of solutions to be randomly generated. It belongs to the class of meta-heuristic approaches, which can be potentially applied to any optimization problem. As for all meta-heuristic searches, SA can give no guarantees to converge towards the global optimum.

In 1988, SA was proposed [46] as a valid alternative to the MMDG method outlined in Sect. 16.7. The MMDG, actually, has been reduced to a preprocessing step for generating initial candidate solutions to be given to SA. The employment of SA overcame some important issues, such as the one of finding embeddings in spaces with a dimension higher than 3 (see Sect. 16.7). In SA, indeed, the solution space is fixed and represented by a subregion of the Euclidean three-dimensional space. Shortly afterwards, the MMDG step was abandoned altogether [50].

To date, this is the method that is mostly used for the determination of protein conformations from NMR data (as a quick search on the PDB [2] can show). This can be due to the ease of implementation of meta-heuristics such as SA, as well as to the availability of software tools where SA is implemented together with other useful tools for managing NMR data. Available software include ARIA [48], CYANA [16], and UNIO [15]. In these implementations, SA represents one step of a more complex procedure, where, for example, ambiguous NOEs (see Sect. 16.4) are verified by exploiting partial solutions found by SA.

The whole procedure, however, and in particular the SA-based step, is heuristic. Decisions taken during the procedure, such as random modifications in candidate solutions or the rejection of some ambiguous distances on the basis of partially obtained solutions, can lead the search in the wrong direction, without any possibility to backtrack. It is important to remark that, even if the procedure can identify a solution for which all available distances are satisfied, this does not imply that it represents the actual protein structure. All possible conformations should be identified, and the ones having the most evident biological sense should be taken into consideration.

16.8.2 A Discrete and Exact Method for the MDGP

In the formulated global optimization problem, the domain of the penalty function (such as, e.g., the function (16.4)) generally corresponds to a subregion of the three-dimensional Euclidean space. As a consequence, an infinite number of potential solutions are contained in this subregion, because it is continuous. The SA approach discussed in the previous section is based on a search in such a continuous space. Under certain assumptions, however, this subregion can be transformed in a discrete domain, where a finite number of potential solutions is contained.

Let G be a weighted undirected graph representing an instance of the MDGP with exact distances. As discussed in Sect. 16.5, if G is a trilateration graph, then there is information enough to solve the problem in polynomial time. For a given ordering on its vertex set, a trilateration graph is such that, for each $v \in V$ with $v > 3$, there are at least four vertices $u < v$ such that the distances between any u and v is known. We say that, in this case, there are four *reference distances* for the vertex v. In this hypothesis, the only feasible position for this vertex can be computed by intersecting the four spheres defined by the four available distances regarding v

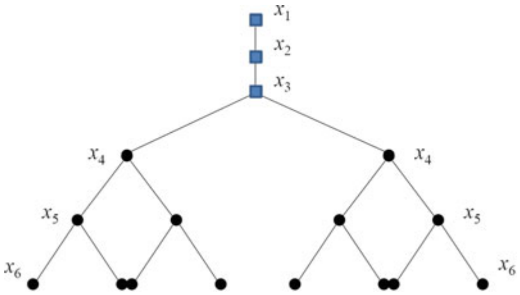

Fig. 16.1 The search domain of a discretizable MDGP instance with exact distances

(see Sect. 16.5). By iterating this procedure from the vertex $v = 4$ until the last one in the ordering associated to the graph G, the molecular conformation can be obtained in only $|V| - 3$ steps.

This is an extreme case allowing for discretization. Instead of an infinite number of positions belonging to a continuous space, there is only one possible atomic position for each $v \in V$. The discretization is however still possible when weaker assumptions are satisfied [29]. If, for each $v > 3$, at least three (not four) vertices u are available so that the distances between each u and v is known, then two possible positions for v (not only one) can be computed by intersecting three spheres (see Sect. 16.5). In this case, the MDGP cannot be solved in polynomial time, because the new search domain is a binary tree organized in n layers, each one containing the possible coordinates of a certain vertex in G (see Fig. 16.1). On the last layer, there are 2^{n-3} possible atomic positions for the last vertex in the order associated to the graph G. As a consequence, this tree contains 2^{n-3} potential solutions to the MDGP.

On the basis of the consecutivity assumption for the reference distances, there are two classes of discretizable MDGP instances that can be defined. In the DMDGP [29], for each vertex $v > 3$, the three reference distances are between v and, respectively, $v-1$, $v-2$, and $v-3$. In the DDGP [43], the reference distances can refer to any vertex which is smaller than v in rank. As a consequence, it can be proved that the class of DMDGP instances is contained in the DDGP class. In the following, we will not make a precise distinction between the DMDGP and the DDGP. Therefore, we will say in general that an instance is *discretizable* if it belongs to one of these two subclasses of the MDGP.

The Branch and Prune (BP) algorithm [33] is based on the idea of efficiently exploring discrete search domains. It can be applied only in case the discretization assumptions are satisfied. The basic idea is to construct the binary tree step by step, i.e., atomic position by atomic position, and to verify the feasibility of such atomic positions as soon as they are computed. Suppose that a partial set of coordinates has already been computed for the vertices $u \in V$ which are smaller in rank than a certain given vertex v. By intersecting the three spheres as explained before, we can obtain the two corresponding possible positions for v. Then, by using some additional information on distances regarding v that was not employed in the sphere intersection, the feasibility of such atomic positions can be verified. In case the position is not feasible (it does not satisfy at least one of the available distance constraints), then it can be removed from the tree. Moreover, the whole tree branch starting from this position can be pruned as well. The pruning phase of BP is its strong point.

Differently from meta-heuristic searches (such as the SA-based algorithm in Sect. 16.8.1) and methods which are based on an exploration of a continuous domain, the BP algorithm is a deterministic algorithm, which is potentially able to enumerate all solutions for a given instance of the MDGP. This point is crucial in protein structure determination. All possible conformations for a certain set of distances should be computed and successively analyzed.

As discussed in Sect. 16.4, NMR instances of the MDGP mostly contain interval distances (and not exact distances). In this case, however, even if the complexity of the problem increases with the uncertainty associated to the interval distances, the discretization is still possible. Let us suppose, for example, that, for a certain $v \in V$, two reference distances are exact, while the third distance is represented by an interval. In the sphere intersection, therefore, one of the spheres needs to be replaced by a spherical shell so that the new intersection consists, most likely, of two disjoint curves. In order to guarantee the discretization, a certain number of sample distances must be taken from the available interval, and a predetermined number of possible atomic positions on the two curves needs to be selected [32].

An instance of the MDGP is represented by a weighted undirected graph G and by a vertex order for the vertices in G. Since the assumptions for the discretization strongly depend upon the given vertex order, changing the order can transform an MDGP instance into a discretizable instance, and vice versa. Therefore, given a graph G, it is interesting to verify whether there exist vertex orders that allow for the discretization [28].

This task is more complex when there are distances that are represented by intervals. In addition to the requirement on the presence of the distances necessary for performing the discretization, other conditions on the distance type (exact distance or interval) may need to be considered. When only one reference distance is an interval, indeed, the discretization can be performed by applying the strategy mentioned above (the intersection among two spheres and one spherical shell). When more than one reference distance is an interval, the intersection can give more complex Euclidean objects so that the definition of vertex orders avoiding for their generation can be necessary.

In [31, 32], instead of using an automatic tool, a vertex order has been handcrafted which allows for discretizing MDGPs concerning protein backbones and containing NMR data (see Fig. 16.2). This order is constructed so that, for each vertex $v > 3$, only one reference distance related to v can be represented by an interval, whereas the other two are always exact. Similar orders for some protein side chains have been proposed in [6]. All these orders exploit distances derived from the chemical composition of proteins for the discretization process, whereas NMR distances are only employed for pruning purposes. This way, the discrete search domain cannot be affected by errors due to the NMR experiments. In order to

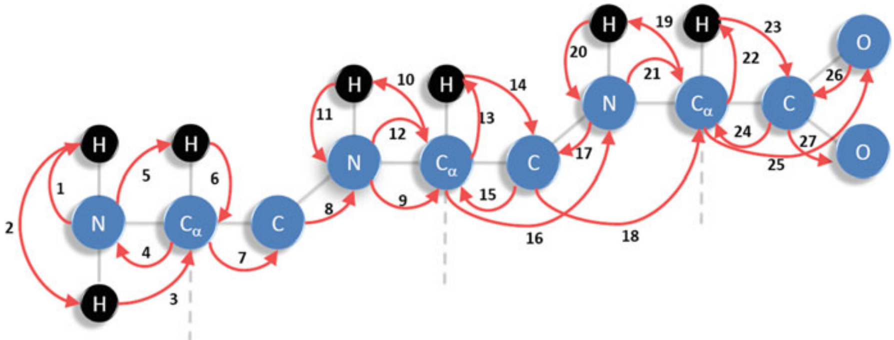

Fig. 16.2 The handcrafted order for the discretization of protein backbones

consider NMR distances for pruning purposes only, cycling is possible in the orders, i.e., the same atom can be represented by more than one vertex of the graph G.

Vertex orders can also be generated so that the maximum width of the corresponding trees can be controlled. If pruning is performed by exploiting NMR distances (as in these hand-crafted orders), then it is important to place the hydrogen atoms (see Sect. 16.4) in strategic positions: if they are too far from each other in the order, pruning is not possible on too many consecutive layers of the tree, allowing for a consistent combinatorial explosion. A deep study on the width of BP trees can be found in [35] in the case all distances are exact.

The *interval* BP (*i*BP) [32] is an extension of the BP algorithm that is able to manage interval data. It has been conceived in order to manage the three following situations. First, the current vertex refers to a duplicated atom, i.e., to an atom which was already considered earlier in the order. In this case, the algorithm simply assigns to this vertex the same position of its previous copy (this implies that cycling does not increase the complexity of the problem). Second, the three reference distances for the current vertex are all exact, and the sphere intersection provides the only two possible positions. Third, one of the reference distances is represented by an interval. In this case, we need to intersect two spheres with a spherical shell, and this intersection provides two curves in the three-dimensional space. In order to discretize, we choose D sample distances from the interval, and we intersect the corresponding three spheres D times. As a consequence, $2 \times D$ possible atomic positions are determined for the current vertex.

Another important point in BP is the fact that it can manage wrongly assigned distances [39] in a deterministic way. Instead of pruning a tree branch as soon as an atomic position does not satisfy one of the distance constraints, the idea is to delay the pruning phase until a predefined number of violations are found. This approach, unfortunately, can be inefficient when the predefined maximum number of violations is large enough to significantly increase the tree width. Work is currently in progress for overcoming this issue.

16.9 New Perspectives in NMR Distance Geometry

Discovering the three-dimensional structure of molecules such as proteins is a very important and challenging problem in biology and biomedicine. In recent years, the research community actively worked on this problem, known as the MDGP. Despite this great effort, a lot of research still need to be performed in order to identify good-quality conformations of biological molecules.

The solution to an MDGP from NMR data can be mainly divided in two main steps. Firstly, the molecule is isolated and analyzed in solution by NMR spectroscopy (see Sect. 16.3); then, the distance information provided by the experiments is exploited for the construction of the molecular conformation (see Sect. 16.5). Both steps are strongly multidisciplinary so that biologists, chemists, physicists, mathematicians, and computer scientists can work in concert on efficient and reliable solution methods.

In spite of this fact, there are nowadays not so many interactions among these communities. In the biological community, the currently used methods for the solution of MDGPs containing NMR data are all based on the meta-heuristic SA (see Sect. 16.8.1), which can give no guarantees of optimality. On the other side, the operational research community developed several more sophisticated and accurate methods for the MDGP (see [30,34,36] for recent surveys). However, some methods rely on assumptions that may not be satisfied in biology, or their performances have never been evaluated on real NMR data. It is worth remarking that the SA-based methods for MDGP would not be able to provide any approximation to solutions if they were not coupled with appropriate tools for NMR management.

The BP algorithm (see Sect. 16.8.2) is a recent algorithm for the MDGP whose development is performed in a strong multidisciplinary collaboration. Firstly developed for solving artificial MDGP instances [33], the algorithm has been adapted successively for solving NMR instances [41]. It is very promising because of its deterministic nature. Differently from SA-based methods for the MDGP, the BP algorithm is potentially able to identify all the conformations satisfying the distance constraints. In other words, BP can enumerate all solutions to the mathematical problem, to be filtered later in order to discovering the most probable biological conformations. Any other conformation which is not contained in the BP solution set cannot be a solution to the problem (this statement is not true when meta-heuristic methods are employed). The development of BP is currently in progress and we believe it could be a valid alternative to currently employed methods.

Another interesting point for future research is the following. As discussed in Sect. 16.4, there is actually another problem to be solved prior the formulation of an MDGP with NMR data: the NOE assignment problem. In some cases this problem can be tough, especially in presence of ambiguous NOEs, and it can be the source of errors. We foresee therefore the possibility of integrating this assignment problem inside the BP algorithm. The employment of a completely deterministic method could allow for overcoming many of the issues causing errors in other methods.

Acknowledgements This work is partially supported by the ANR project ANR-10-BINF-03-01, “Bip:Bip”. TM and MN acknowledge support by the CNRS and the Institut Pasteur.

References

1. Almeida, F., Moraes, A., Neto, F.G.: Overview on protein structure determination by NMR — Historical and future perspectives of the use of distance geometry. In: Mucherino et al, “Distance Geometry: Theory, Methods, and Applications”
2. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acid Res. **28**, 235–242 (2000)
3. Blumenthal, L.M.: Theory and Application of Distance Geometry. Chelsea, New York (1970)
4. Braun, W., Bösch, C., Brown, L.R., Gō, N., Wüthrich, K.: Combined use of proton-proton Overhauser enhancements and a distance geometry algorithm for determination of polypeptide conformations. Application to micelle-bound glucagon. Biochimica et Biophysica Acta **667**, 377–396 (1981)
5. Clore, G.M., Gronenborn, A.M.: Determination of three-dimensional structures of proteins and nucleic acids in solution by nuclear magnetic resonance spectroscopy. Crit. Rev. Biochem. Mol. Biol. **24**, 479–564 (1989)
6. Costa, V., Mucherino, A., Lavor, C., Carvalho, L.M., Maculan, N.: On suitable orders for discretizing molecular distance geometry problems related to protein side chains. In: IEEE Conference Proceedings, Federated Conference on Computer Science and Information Systems (FedCSIS12), Workshop on Computational Optimization (WCO12), Wroclaw, Poland, September 9–12 (2012)
7. Crippen, G.M.: A novel approach to calculation of conformation: distance geometry. J. Comput. Phys. **24**(1), 96–107 (1977)
8. Crippen, G.M., Havel, T.F.: Stable calculation of coordinates from distance information. Acta Crystallographica Section A**34**, 282–284 (1978)
9. Crippen, G.M., Havel, T.F.: Distance Geometry and Molecular Conformation. Wiley, New York (1988)
10. Davis, R.T., Ernst, C., Wu, D.: Protein structure determination via an efficient geometric build-up algorithm. BMC Struct. Biol. **10**:S7 (2010)
11. Easthope, P.L., Havel, T.F.: Computational experience with an algorithm for tetrangle inequality bound smoothing. Bull. Math. Biol. **51**, 173–194 (1991)
12. Engh, R.A., Huber, R.: Accurate bond and angle parameters for X-ray structure refinement. Acta Crystallographica Section A **47**, 392–400 (1991)
13. Eren, T., Goldenberg, D.K., Whiteley, W., Yang, Y.R., Morse, A.S., Anderson, B.D.O., Belhumeur, P.N.: Rigidity, computation, and randomization in network localization. In: IEEE Infocom Proceedings, 2673–2684 (2004)
14. Grishaev, A., Llinas, M.: Protein structure elucidation from NMR proton densities. Proc. Nat. Acad. Sci. USA **99**, 6713–6718 (2002)
15. Guerry, P., Herrmann, T.: Comprehensive automation for NMR structure determination of proteins. Meth. Mol. Biol. **831**, 33–56 (1992)
16. Güntert, P.: Automated NMR structure calculation with CYANA. In: Downing, A.K. (ed.) Protein NMR Techniques. Meth. Mol. Biol. **278**, 353–378 (2004)
17. Güntert, P., Berndt, K.D., Wüthrich, K.: The program ASNO for computer-supported collection of NOE upper distance constraints as input for protein structure determination. J. Biomol. NMR **3**, 601–606 (1993)
18. Havel, T.F.: Distance geometry. In: Grant, D.M., Harris, R.K. (eds.) Encyclopedia of Nuclear Magnetic Resonance, pp. 1701–1710. Wiley, New York (1995)
19. Havel, T.F., Kunts, I.D., Crippen, G.M.: The theory and practice of distance geometry. Bull. Math. Biol. **45**, 665–720 (1983)

20. Horst, R., Pardalos, P.M.: Handbook of Global Optimization. Springer (1994), http://www.springer.com/mathematics/book/978-0-7923-3120-9
21. Huang, A., Stultz, C.M.: Finding order within disorder: elucidating the structure of proteins associated with neurodegenerative disease. Future Med. Chem. **1**, 467–482 (2009)
22. Janin, J.: Protein-protein docking tested in blind predictions: the CAPRI experiment. Mol. Biosyst. **6**, 2351–2362 (2010)
23. Kirkpatrick, S., Jr. Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science **220**(4598), 671–680 (1983)
24. Kleywegt, G.J.: Validation of protein models from C_α coordinates alone. J. Mol. Biol. **273**(2), 371–376 (1997)
25. Kline, A.D., Braun, W., Wüthrich, K.: Studies by ^{1}H nuclear magnetic resonance and distance geometry of the solution conformation of the a-amylase inhibitor Tendamistat. J. Mol. Biol. **189**, 377–382 (1986)
26. Koning, T.M., Davies, R.J., Kaptein, R.: The solution structure of the intramolecular photoproduct of d(TpA) derived with the use of NMR and a combination of distance geometry and molecular dynamics. Nucleic Acids Res. **18**, 277–284 (1990)
27. Kuszewski, J., Nilges, M., Brünger, A.T.: Sampling and efficiency of metric matrix distance geometry: a novel partial metrization algorithm. J. Biomol. NMR **2**, 33–56 (1992)
28. Lavor, C., Lee, J., Lee-St.John, A., Liberti, L., Mucherino, A., Sviridenko, M.: Discretization orders for distance geometry problems. Optim. Lett. **6**(4), 783–796 (2012)
29. Lavor, C., Liberti, L., Maculan, N., Mucherino, A.: The discretizable molecular distance geometry problem. Comput. Optim. Appl. **52**, 115–146 (2012)
30. 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)
31. Lavor, C., Liberti, L., Mucherino, A.: On the solution of molecular distance geometry problems with interval data. In: IEEE Conference Proceedings, International Workshop on Computational Proteomics (IWCP10), International Conference on Bioinformatics & Biomedicine (BIBM10), Hong Kong, 77–82 (2010)
32. Lavor, C., Liberti, L., Mucherino, A.: The *interval* Branch-and-Prune algorithm for the discretizable molecular distance geometry problem with inexact distances, to appear in J. Global Optim. (2012), http://link.springer.com/article/10.1007%2Fs10898-011-9799-6
33. 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)
34. Liberti, L., Lavor, C., Mucherino, A., Maculan, N.: Molecular Distance Geometry Methods: from continuous to discrete. Int. Trans. Oper. Res. **18**, 33–51 (2010)
35. Liberti, L., Masson, B., Lavor, C., Mucherino, A.: Branch-and-Prune trees with bounded width. In: Proceedings of the 10th Cologne-Twente Workshop on Graphs and Combinatorial Optimization (CTW11), Rome, Italy, 189–193 (2011)
36. Liberti, L., Lavor, C., Maculan, N., Mucherino, A.: Euclidean distance geometry and applications, Tech. Rep. 1205.0349v1 [q-bio.QM], arXiv (2012)
37. Mucherino, A., Costantini, S., di Serafino, D., D'Apuzzo, M., Facchiano, A., Colonna, G.: Towards a Computational Description of the Structure of all-alpha Proteins as Emergent Behaviour of a Complex System. Comput. Biol. Chem. **32**(4), 233–239 (2008)
38. Mucherino, A., Seref, O.: Modeling and solving real life global optimization problems with meta-heuristic methods. In: Papajorgji, P.J., Pardalos, P.M. (eds.) Advances in Modeling Agricultural Systems, pp. 403–420 (2008) http://link.springer.com/chapter/10.1007%2F978-0-387-75181-8_19
39. Mucherino, A., Liberti, L., Lavor, C., Maculan, N.: Comparisons between an exact and a meta-heuristic algorithm for the molecular distance geometry problem. In: ACM Conference Proceedings, Genetic and Evolutionary Computation Conference (GECCO09), Montréal, Canada, 333–340 (2009)
40. Mucherino, A., Papajorgji, P., Pardalos, P.M.: Data Mining in Agriculture. Springer, New York (2009)

41. Mucherino, A., Lavor, C., Malliavin, T., Liberti, L., Nilges, M., Maculan, N.: Influence of pruning devices on the solution of molecular distance geometry problems. In: Pardalos, P.M., Rebennack, S. (eds.) Lecture Notes in Computer Science **6630**, Proceedings of the 10th International Symposium on Experimental Algorithms (SEA11), Crete, Greece, 206–217 (2011)
42. Mucherino, A., Lavor, C., Liberti, L.: Exploiting symmetry properties of the discretizable molecular distance geometry problem. J. Bioinformatics Comput. Biol. **10**(3), 1242009 (2012)
43. Mucherino, A., Lavor, C., Liberti, L.: The discretizable distance geometry problem, Optim. Lett. **6**(8), 1671–1686 (2012)
44. Nazareth, J.L.: Conjugate gradient method. Wiley Interdiscipl. Rev. Comput. Stat. **3**(1), 348–353 (2009)
45. Nilges, M.: Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. J. Mol. Biol. **245**, 645–660 (1995)
46. Nilges, M., Clore, G.M., Gronenborn, A.M.: Determination of three-Dimensional structures of proteins from interproton distance data by hybrid distance geometry – dynamical simulated annealing calculations. Fed. Eur. Biochem. Soc. **229**, 317–324 (1988)
47. Nilges, M., Marcias, M.J., O'Donoghue, S.I.: Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from β-spectrin. J. Mol. Biol. **269**, 408–422 (1997)
48. Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T.E., Nilges, M.: ARIA2: Automated NOE assignment and data integration in NMR structure calculations. Bioinformatics **23**(3), 381–382 (2007)
49. Sali, A., Blundell, T.L.: Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. **234**, 779–815 (1993)
50. Saxe, J.B.: 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)
51. Shen, Y., Delaglio, F., Cornilescu, G., Bax, A.: TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J. Biomol. NMR **44**, 213–223 (2009)
52. Spedding, M.: Resolution of controversies in drug/receptor interactions by protein structure. Limitations and pharmacological solutions. Neuropharmacology **60**, 3–6 (2011)
53. Weber, P.L., Morrison, R., Hare, D.: Determining stereo-specific ^{1}H nuclear magnetic resonance assignments from distance geometry calculations. J. Mol. Biol. **204**, 483–487 (1988)
54. Wishart, D.S., Sykes, B.D.: The ^{1}3C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. J. Biomol. NMR **4**, 171–180 (1994)
55. Wu, D., Wu, Z.: An updated geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data. J. Global Optim. **37**, 661–673 (2007)
56. Wüthrich, K., Billeter, M., Braun, W.: Pseudo-structures for the 20 common amino acids for use in studies of protein conformations by measurements of intramolecular proton-proton distance constraints with nuclear magnetic resonance. J. Mol. Biol. **169**(4), 949–961 (1983)