Minimal NMR distance information for rigidity of protein graphs

Carlile Lavor, Leo Liberti, Bruce Donald, Thérèse Malliavin, Michael Nilges

Minimal NMR distance information for rigidity of protein graphs

Carlile Lavor[a,*], Leo Liberti[b], Bruce Donald[c], Thérèse E. Malliavin[d], Michael Nilges[e]

[a] *University of Campinas (IMECC-UNICAMP), 13081-970, Campinas-SP, Brazil*
[b] *CNRS LIX, École Polytechnique, 91128 Palaiseau, France*
[c] *Duke University, Department of Computer Science, Durham, NC 27708-0129 USA*
[d] *Institute Pasteur and CNRS, Unité de Bioinformatique Structurale, Paris 75015, France*
[e] *Institute Pasteur, Unité de Bioinformatique Structurale, Paris 75015, France*

Abstract

Nuclear Magnetic Resonance (NMR) experiments provide distances between close atoms of a protein molecule and the problem is how to determine the 3D protein structure by exploiting such distances. We present a new hand-crafted order on the atoms of the protein that uses information from the chemistry of proteins and NMR experiments and allows us to formulate the problem as a combinatorial search. Additionally, this order tell us what kind of NMR distance information is crucial to understand the cardinality of the solution set of the problem and its computational complexity.

Keywords: `elsarticle.cls`, LaTeX, Elsevier, template

2010 MSC: 00-01, 99-00

1. Protein Structure and Distance Geometry

The 3D protein structure determination is of fundamental importance for studying the protein function [18]. Indeed, biochemical mechanisms taking place in protein structure are the basic steps hidden behind all biological processes, as cell division, protein translation, invasion of host cells by pathogens, and

*Corresponding author
Email addresses: `clavor@ime.unicamp.br` (Carlile Lavor), `liberti@lix.polytechnique.fr` (Leo Liberti), `brd@cs.duke.edu` (Bruce Donald), `terez@pasteur.fr` (Thérèse E. Malliavin), `michael.nilges@pasteur.fr` (Michael Nilges)

communication between cells. In that way, the determination of protein structures can be considered as the building of a bridge between the description of biological cellular processes and the world of physico-chemistry.

At the first days of structural biology, in the fifties, the X-ray crystallography was the first method allowing the determination of protein structure. Therefore, the crystallized proteins were perceived as rigid objects, displaying mostly a unique conformation, with more of less large harmonic vibrations around this conformation. The development of Nuclear Magnetic Resonance (NMR) permitted, starting from the nineties, to study proteins structure in solution. The development of NMR relaxation methods put in evidence the internal dynamics in the protein structures and gave a more flexible vision of them [26].

The protein internal flexibility was then recognized as playing a very important role into many biological processes. The intrinsically disordered proteins are thought to be functional important proteins, even if they lack a precisely defined 3D structure. The misfolded proteins, undergoing a conversion from their native structure into amyloid, induce the development of neurodegenerative diseases. The allostery phenomenon, allowing long-range communication through the protein structure, is thought to be involved into any activity mechanism of proteins.

The NMR structure determination is mainly based on the measurement of inter-atomic distances, determined through the observation of the Nuclear Overhauser Effect (NOE), which is induced by the transfer of magnetization through dipolar coupling between the observed hydrogens. The obtained distance values are quite qualitative, due to the numerous paths for the magnetization transfer and to the molecular internal dynamics. The corresponding distance restraints are thus usually applied as interval restraints.

NMR experiments can be used to determine some (short) Euclidean distances between hydrogen atoms in a protein and the problem is to determine its 3D structure based on this partial distance information.

We can use two types of sets (the entities V and their relationships E) and a real function d on E to model this problem: V represents the set of atoms, E

represents the set of atom pairs for which a distance is available, and the function $d : E \to [0, \infty)$ assigns nonnegative real numbers to each pair $\{u, v\} \in E$ (the fact that we allow a distance to be zero will be explained in Section 5).

When we consider V, E, d together, we have a *graph*, denoted by $G = (V, E, d)$. We say that G is a weighted simple undirected graph because we associate values (distances) to the elements of E, if $\{u, v\} \in E$ then $u \neq v$, and $\{u, v\} = \{v, u\}$.

The way to represent a molecule consisting of a set of atomic symbols linked by segments was originally described in [17] and, in fact, the origin of the word graph is due to the representation of molecules [60]. This relationship between molecules and graphs is probably the deepest existing between chemistry and discrete mathematics.

A graph $G = (V, E, d)$ is just a mathematical abstraction to represent the problem data. The problem itself is to find a function $x : V \to \mathbb{R}^3$ that associates each element of V with a point in $\mathbb{R}^3$ in such a way that the Euclidean distances between the points correspond to the values given by d. This is a *Distance Geometry Problem* (DGP) in $\mathbb{R}^3$, formally described as follows.

Definition 1. *Given an integer $K > 0$ and a simple undirected graph $G = (V, E, d)$ whose edges are weighted by a function $d : E \to [0, \infty)$, find a function $x \colon V \to \mathbb{R}^K$ such that*

$$\forall \{u, v\} \in E, \ ||x_u - x_v|| = d_{uv}, \tag{1}$$

where $x_u = x(u)$, $x_v = x(v)$, $d_{uv} = d(\{u, v\})$, and $||x_u - x_v||$ is the Euclidean distance between x_u and x_v.

From now on, we will fix $K = 3$, since we are interested in the application of the DGP to protein conformation [16]. Recent surveys on Distance Geometry (DG) are given in [7, 45], an edited book with different applications can be found in [50], and some DG historical notes are presented in [46].

In 1983, the first distance geometry-based method for molecular conformation was proposed [27] and, in 1984, the first protein structure was determined

in its native solution state from NMR data [28].

The simplest approach to the problem is to directly try to solve the set of equations (1). However, there is evidence that a closed-form solution is not possible [5]. Since those equations also pose difficulties to be solved numerically, a common approach is to formulate the DGP as a nonlinear global minimization problem,

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^3} \sum_{\{u,v\} \in E} (||x_u - x_v||^2 - d_{uv}^2)^2,$$

where $|V| = n$. However, solving such a problem is hard from a computational complexity point of view, as well as from a practical one [45, 58, 59]. In [36], some global optimization algorithms have been tested but none of them scale well to medium or large instances. A survey on different methods to the DGP is given in [43].

Assuming the input data are correct and precise (see Section 5 for other cases), the set X of solutions of a DGP will yield all the 3D structures of the protein that are compatible with the given distances. Any $x \in X$ can be translated and rotated in $\mathbb{R}^3$ implying that the solution set is not only infinite, but uncountable. However, if we do not consider the effect of translations and rotations, the cardinality of X depends on the structure of the associated graph $G = (V, E, d)$. If the set of edges E contains all possible pairs from V, there is only one solution which can be found in linear time [19]. In general, the problem is NP-hard [55].

From algebraic geometry, it is possible to prove that there are just two possibilities regarding the cardinality of the solution set X: it is either finite or uncountable, supposing that $X \neq \varnothing$ [6]. This result is strongly related to graph rigidity: almost all nonrigid graphs yield uncountable cardinalities for X [25].

If the graph is rigid, the solution set is finite (up to translations and rotations). In this case, a combinatorial search is better suited than a continuous one, because in addition to the accuracy and efficiency of combinatorial methods, the graph rigidity allows us to get more information about the cardinality and the structure of the solution set X [40, 47].

The original contribution of this paper is theoretical, presenting a new hand-crafted order on the vertices of the protein DGP graph G that uses information from the chemistry of proteins and NMR experiments. This order guarantees the rigidity of G and, most importantly, "organizes the search space" in such a way that it can be searched efficiently for all the mathematical solutions of the problem. Also, it tell us what kind of information from the NMR experiments is crucial to understand the cardinality of the solution set and the computational complexity of the problem.

To explain the properties of the proposed order, important connections among NMR protein structure, distance geometry, graph rigidity, and graph vertex orders are established. We tried to do that without excessive formalism, although all the important concepts and results from those research areas are presented.

In the next section, the main results from graph rigidity necessary here are given. Section 3 shows the importance of vertex orders in DGP graphs. Section 4 presents the discrete version of the DGP. In Section 5, the new hand-crafted order is defined with its most important properties. Finally, we end with conclusions and some new research directions in Section 6.

2. Graph Rigidity and Distance Geometry

Given a graph $G = (V, E, d)$ of a DGP, a function $x\colon V \to \mathbb{R}^3$ is called a *realization* of the graph in $\mathbb{R}^3$. If x satisfies all the equations (1), it is a *valid realization*. The pair G with a valid realization x is a *framework* (G, x).

In order to use frameworks to model protein structures and to have precisely the notion of rigidity of a framework [31], we need to define two relations: isometry and congruency.

Two frameworks (G, x) and (G, y) are *isometric*, denoted by $(G, x) \sim (G, y)$, if

$$\forall \{u, v\} \in E,\ ||x_u - x_v|| = ||y_u - y_v||,$$

and *congruent*, denoted by $(G, x) \equiv (G, y)$, if

$$\forall u, v \in V,\ ||x_u - x_v|| = ||y_u - y_v||.$$

Thus, two frameworks are congruent only if all pairs of vertices from V have the same related distances, not only the pairs in E. Trivially, congruency implies isometry, but the converse is not true in general. We remark that any congruence is a composition of translations, rotations, and reflections [8].

(G, x) is a *rigid framework* if there exists a real number $\varepsilon > 0$ such that

$$(G, y) \sim (G, x) \text{ and } \parallel x_v - y_v \parallel < \varepsilon,\ \forall v \in V \Rightarrow (G, y) \equiv (G, x).$$

Geometrically, this means that a framework is rigid if it has no continuous deformations aside from composition of translations, rotations and reflections. That is, the only way to continuously move a point in a rigid framework is moving all points such that all pairwise distances are preserved, and not only those given by the edges. Using the concept of infinitesimal rigidity of a framework [61], we can define graph rigidity.

Let (G, x) be a framework in $\mathbb{R}^3$, where $|V| = n$ and $|E| = m$. Consider the linear system $R\lambda = 0$, where $\lambda \in \mathbb{R}^{3n}$ and R is the $m \times 3n$ matrix, each $\{u, v\}$th row of which has exactly 6 nonzero entries given by

$$x_i(u) - x_i(v) \text{ and } x_i(v) - x_i(u),\ \{u, v\} \in E \text{ and } i = 1, 2, 3,$$

where $x_1(u), x_2(u), x_3(u)$ are the Cartesian coordinates of x_u in $\mathbb{R}^3$.

The framework is *infinitesimally rigid* if the only solutions of $R\lambda = 0$ are translations or rotations. Infinitesimal rigidity implies rigidity [22], and if a graph has a unique infinitesimally rigid framework, then almost all its frameworks are rigid [29].

It makes sense then to define a *rigid graph* as a graph having an infinitesimally rigid framework. There is also a notion of a graph being rigid independently of the framework assigned to it, known as *generic rigidity* [14], that will not be used here.

A characterization of all rigid graphs in $\mathbb{R}^2$ was described by Laman [34], but no such complete characterization is known in $\mathbb{R}^3$. A heuristic method was introduced in [57] and current conjectures can be found in [32].

If a DGP graph has a unique valid realization, up to congruences, it is called *globally rigid*. In [14], necessary and sufficient conditions for global rigidity in $\mathbb{R}^2$ were presented. Hendrickson [29] conjectured that the same conditions would be sufficient for $\mathbb{R}^3$, but this was disproved by Connelly [14]. Some graph properties ensuring global rigidity in $\mathbb{R}^2$ and $\mathbb{R}^3$ are given in [4].

3. Vertex Orders and Distance Geometry

The idea of exploiting vertex order to investigate graph rigidity first appeared in [30]. In fact, vertex orders are important to solve many problems modeled by graphs [9, 51].

If there is a *trilateration order* in a DGP graph (every vertex beyond the first 4 is adjacent to at least 4 predecessors), it is globally rigid and such order makes it possible to triangulate the position of each next vertex. This implies a *linear time* algorithm to find the *unique* incongruent solution [20].

Adjacent predecessors in a vertex order are critical: any fewer, and the number of DGP solutions might be uncountable; any more, and the corresponding DGP can be solved uniquely in linear time [45]. So, the number of adjacent predecessors, in a given order, is related to the cardinality of the DGP solution set and also to the required computational effort to find a solution.

In general, we do not have trilateration orders in protein graphs $G = (V, E, d)$ [39], but using the information provided by NMR experiments and chemistry of proteins, we can try to find vertex orders $v_1, ..., v_n \in V$ such that:

- The first 3 vertices form a *clique*:

$$\{v_1, v_2\}, \{v_1, v_3\}, \{v_2, v_3\} \in E;$$

- Each vertex with rank greater than 3 is adjacent to at least 3 predecessors:

$$\forall i > 3, \exists\ j, k, l \text{ with } j < i, k < i, l < i : \{v_j, v_i\}, \{v_k, v_i\}, \{v_l, v_i\} \in E.$$

The class of DGP instances possessing these orders, where the initial clique has a valid realization and the strict triangular inequalities related to the adjacent predecessors v_j, v_k, v_l to v_i, $i > 3$, are satisfied (*i.e.* $d_{v_j v_k} + d_{v_k v_l} > d_{v_j v_l}$), is called the *Discretizable Distance Geometry Problem* (DDGP), and the orders themselves DDGP orders [23, 48].

The initial clique guarantees that the solution set X will contain just incongruent solutions and strictness of the triangular inequality prevents an uncountable quantity of solutions [48]. In the same paper, it was proved that the graph of any DDGP instance is rigid and an exact solution method, called Branch-and-Prune (BP), was presented for finding all incongruent solutions. BP can be exponential in the worst case, since the DDGP is an NP-hard problem [10, 42, 48].

In a DDGP order, the fourth vertex v_4 can be realized solving the quadratic system (to simplify the notation, we will use x_i instead of x_{v_i} and $d_{i,j}$ instead of $d_{v_i v_j}$)

$$\begin{aligned} ||x_4 - x_1||^2 &= d_{1,4}^2 \\ ||x_4 - x_2||^2 &= d_{2,4}^2 \\ ||x_4 - x_3||^2 &= d_{3,4}^2, \end{aligned}$$

which can result in up to two possible positions for v_4. Using the same strategy, for each position already determined for v_4, we obtain other two for v_5, and so on. Because of the rigidity of the DDGP graph, the search space is finite, having 2^{n-3} possible solutions.

If we have an "extra" distance information, $\{v_r, v_i\} \in E$ with $r < i$, we can add more one equation to the system related to $v_i, i > 3$, resulting in

$$\begin{aligned} ||x_i - x_j|| &= d_{j,i} \\ ||x_i - x_k|| &= d_{k,i} \\ ||x_i - x_l|| &= d_{l,i} \\ ||x_i - x_r|| &= d_{r,i}. \end{aligned}$$

Squaring both sides of these equations, we get (x_i^T is the transpose of x_i)

$$\begin{aligned} ||x_i||^2 - 2(x_i^T x_j) + ||x_j||^2 &= d_{j,i}^2 \\ ||x_i||^2 - 2(x_i^T x_k) + ||x_k||^2 &= d_{k,i}^2 \\ ||x_i||^2 - 2(x_i^T x_l) + ||x_l||^2 &= d_{l,i}^2 \\ ||x_i||^2 - 2(x_i^T x_r) + ||x_r||^2 &= d_{r,i}^2. \end{aligned}$$

Now, subtracting one of these equations from the others, we eliminate the term $||x_i||^2$ and obtain a linear system in the variable x_i. If the points x_j, x_k, x_l, x_r are not in the same plane, we have a unique solution x_i^* for v_i, supposing $||x_i^* - x_r|| = d_{r,i}$. When there are other adjacent predecessors of v_i beside v_j, v_k, v_l, one or both possible positions for v_i may be infeasible with respect to those additional distances. If both are infeasible, it is necessary to backtrack and repeat the methodology [48].

The DDGP order "organizes" the search space in a *binary tree* and the additional distance information can be used to reduce the search space by pruning infeasible positions in the tree.

The tree begins with the three fixed positions for the initial clique, x_1, x_2, x_3, and at level $i > 3$, the tree contains all (2^{i-3}) possible positions for vertex v_i. The search ends when a *path* from the root ($i = 1$) of the tree to a leaf node ($i = n$) is found by BP algorithm, in such a way that the related positions in $\mathbb{R}^3$ for all the vertices in the order satisfy the DGP equations (1), encoding a valid realization of G. Considering precise input data, the BP performance is impressive from the points of view of both efficiency and reliability [36, 39]. Although the DDGP is NP-hard, a DDGP order can be found in polynomial-time [38].

In the definition of the DDGP, the only requirement on the adjacent predecessors v_j, v_k, v_l to v_i, $i > 3$, is that the associated strict triangular inequality must be satisfied, making the DDGP very general. However, depending on the instance, the distances $d_{j,i}, d_{k,i}, d_{l,i}$ cannot be well-scaled, increasing the incidence of numerical floating point error in solving the related quadratic system

and, in some cases, making its solution impossible [48]. Additionally, when the vertices $\{v_j, v_k, v_l\}$ do not form a clique, the related quadratic system may not have a solution.

Protein graphs provided by NMR experiments have enough information to allow definition of vertex orders involving *immediately contiguous* adjacent predecessors that can avoid those kinds of problems in DDGP instances.

4. The Discretizable Molecular Distance Geometry Problem (DMDGP)

The class of DGP instances that replaces a DDGP order by one with contiguous adjacent predecessors is called the *Discretizable Molecular Distance Geometry Problem* (DMDGP) and the order itself is a DMDGP order [39]. Formally, the DMDGP is defined as follows.

Definition 2. *Given a DGP graph $G = (V, E, d)$ and a vertex order $v_1, ..., v_n$ such that*

- there exists a valid realization for v_1, v_2, v_3 and
- $\forall i > 3$, the set $\{v_{i-3}, v_{i-2}, v_{i-1}, v_i\}$ is a clique with

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

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

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

The fact that the set $\{v_{i-3}, v_{i-2}, v_{i-1}, v_i\}$ is a clique, $\forall i > 3$, eliminates bad distance scaling and guarantees a non-empty solution set for the quadratic system related to the two possible positions for v_i in terms of the positions already determined for $v_{i-3}, v_{i-2}, v_{i-1}$.

In addition to these properties, the distance information in the clique $\{v_{i-3}, v_{i-2}, v_{i-1}, v_i\}$ also allows us to get the following values:

- $d_{1,2}, \ldots, d_{n-1,n}$ (distances associated to consecutive vertices),

- $\theta_{1,3}, \ldots, \theta_{n-2,n}$ (angles in $(0, \pi)$ defined by three consecutive vertices),

- $\cos(\omega_{1,4}), \ldots, \cos(\omega_{n-3,n})$ (cosines of torsion angles in $[0, 2\pi]$ defined by four consecutive vertices), given by [35]:

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

where

$$\begin{aligned} d_{i-3,i-2,i-1} &= d^2_{i-3,i-2} + d^2_{i-2,i-1} - d^2_{i-3,i-1} \\ d_{i-2,i-1,i} &= d^2_{i-2,i-1} + d^2_{i-2,i} - d^2_{i-1,i}. \end{aligned}$$

Using $\cos(\omega_{i-3,i})$, for $i = 4, ..., n$, we obtain two possible values for each torsion angle, implying that we do not need to solve anymore quadratic systems. Computational results presented in [39] demonstrate that avoiding resolution of quadratic systems guarantees more stability in the branching phase of BP.

Another advantage of the DMDGP order is that it is enough to apply the BP (or other algorithm) to find only one solution, since all the others can be easily obtained using *symmetric* properties defined in the BP tree [47, 49]. These properties are also related to the cardinality of the DMDGP solution set, that can be previously calculated based on the DMDGP graph [44].

Considering that the vertex order $v_1, ..., v_n$ represents bonded atoms of a molecule, the values $d_{i-1,i}, \theta_{i-2,i}, \omega_{i-3,i}$ are exactly the *internal coordinates* of the molecule that can also be used to describe its 3D structure [39] (Figure 1).

There is a "price" for all these results: in contrast to DDGP orders, finding a DMDGP order is an NP-complete problem [11], even considering that the initial clique is given. However, exploiting the chemistry of proteins and NMR data, it is possible to design a hand-crafted DMDGP order for any protein graph. We will see that this order can also be used to solve DMDGP instances that incorporate uncertainties from NMR data.

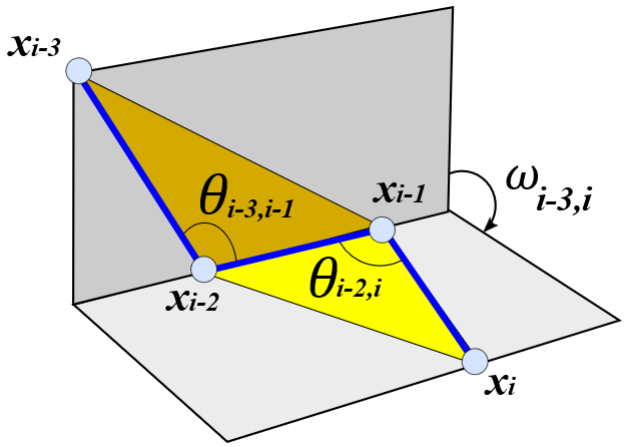

Figure 1: Cartesian and internal coordinates.

5. A New DMDGP Order for Protein Graphs

In order to reduce the number of variables and also the computational effort to solve problems related to protein structure, it is common to assume that all bond lengths and bond angles are fixed at their equilibrium values, which is known as the *Rigid Geometry Hypothesis* [21]. This means that, in terms of internal coordinates, all the values $d_{i-1,i}$, for $i = 2, \ldots, n$, and $\theta_{i-2,i}$, for $i = 3, \ldots, n$, are given *a priori*, and that the 3D protein structure can be determined by the values $\omega_{i-3,i}$, for $i = 4, \ldots, n$. Because of the properties of DMDGP orders, we can also know *a priori* all the values $\cos(\omega_{i-3,i})$, for $i = 4, \ldots, n$, implying that the protein structure is defined by choosing $+$ or $-$ from $\sin(\omega_{i-3,i}) = \pm\sqrt{1 - \cos^2(\omega_{i-3,i})}$, for $i = 4, \ldots, n$. These signs ($+$ or $-$) are obviously related to the branches of the BP tree.

We will consider protein graphs related to the backbone of a protein, the "skeleton" of the molecule, from which its general 3D structure is determined. The protein backbone is a chain of smaller molecules, called amino acids, which are chemically bound to each other. The backbone is defined by a sequence of three atoms, N, C_α, C, where C_α is bound to other group of atoms (the side chains of the protein) that distinguishes one amino acid from another. The

attached atoms to N, C_α, C, respectively H, H_α, O, will be very important to establish our results (Figure 2 presents a backbone with three amino acids). More details about protein graphs including side chains are given in [15, 53, 54].

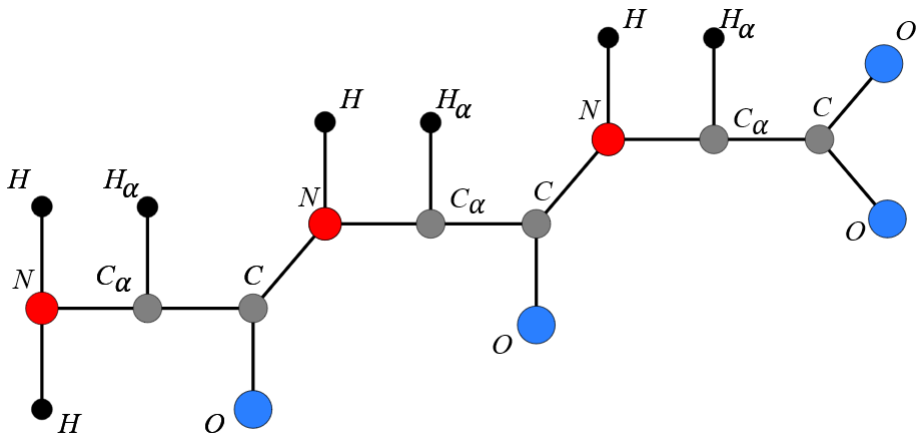

Figure 2: Protein backbone.

5.1. Re-orders

Since we are interested in determining the 3D structure of the backbone of a protein, the sequence of atoms N^i, C_α^i, C^i, for $i = 1, ..., p$ (p is the number of amino acids), would be the first candidate for defining the DMDGP order we are looking for. However, for this kind of order, we do not have all the distances $d_{i-3,i}$ necessary to define a DMDGP instance. By the other hand, since NMR experiments provide distances between hydrogens atoms that are close enough, another order involving only hydrogens could be defined, but this does not work well, mainly because of uncertainty in NMR data [37]. These limitations have been partially addressed by using, at the same time, hydrogens atoms bonded to the backbone and the backbone itself [41].

As done in [41], the idea is to allow the repetition of some vertices in the associated graph, so that at least 3 adjacent predecessors can always be chosen to be contiguous. Such orders are called *re-orders*, defined below. First, the set of

edges E of the protein graph $G = (V, E, d)$ will be partitioned into $E = E' \cup E''$, where $\{u, v\} \in E'$ if $d_{uv} \in (0, \infty)$, and $\{u, v\} \in E''$ if $d_{uv} = [\underline{d}_{uv}, \overline{d}_{uv}]$, with $0 < \underline{d}_{uv} < \overline{d}_{uv}$. Note that the function d is now more general, since some of its values can be intervals that will represent the uncertainties in NMR data. As we will see, E' represents pairs of atoms separated by one and two covalent bonds and E'' represents pairs of hydrogen atoms whose distances are provided by NMR.

Definition 3. *A re-order is a sequence $r : \mathbb{N} \to V \cup \{0\}$, with length $|r| \in \mathbb{N}$ (for which $r_i = r(i) = 0$ for all $i > |r|$), such that*

- $\{r_1, r_2\}, \{r_1, r_3\}, \{r_2, r_3\} \in E'$;
- $\forall i \in \{4, ..., |r|\}$, $\{r_{i-1}, r_i\}, \{r_{i-2}, r_i\} \in E'$;
- $\forall i \in \{4, ..., |r|\}$, $\{r_{i-3}, r_i\} \in E' \cup E''$ *or* $r_{i-3} = r_i$.

The first property says that $d_{r_1 r_2}, d_{r_1 r_3}, d_{r_2 r_3} \in (0, \infty)$ and the second one says that $d_{r_{i-1} r_i}, d_{r_{i-2} r_i} \in (0, \infty)$, for $i = 4, ..., |r|$. That is, all of them must be precise distances and greater than zero.

In the third property, there are 3 possibilities for $d_{r_{i-3} r_i}$, $i = 4, ..., |r|$:

1. $d_{r_{i-3} r_i} = 0$, meaning that there is a vertex repetition ($r_{i-3} = r_i$);
2. $d_{r_{i-3} r_i} \in (0, \infty)$, when $r_{i-3,} r_i$ are related to atoms separated by one of two covalent bonds;
3. $d_{r_{i-3} r_i} = [\underline{d}_{r_{i-3} r_i}, \overline{d}_{r_{i-3} r_i}]$, with $0 < \underline{d}_{r_{i-3} r_i} < \overline{d}_{r_{i-3} r_i}$ (these distances are called *interval distances*).

Requiring $r_i = r_j$, for some $i \neq j$ ($r_{i-3} = r_i$ is a particular case), implies $d_{r_i r_j} = 0$. However, if vertex repetition is used inappropriately, we might end up with a triangle with a side of zero length, which might in turn imply an infinity of possible positions for the next atom (we emphasize the importance of strict triangular inequalities in the definition of the DMDGP). Thus, to preserve discretization, vertex repetition can occur only between pairs $\{r_i, r_j\}$ with $|i - j| \geq 3$. In this case, there is no branching at level $\max\{i, j\}$.

A repetition of a vertex only increases the length of the sequence without affecting the search, since its position in $\mathbb{R}^3$ is already known. However, it can be recomputed in order to control possible numerical instabilities and to check if there are some inconsistencies in the distance information.

To understand what happens when $\{r_{i-3}, r_i\} \in E''$, let us rewrite expression (2) as

$$\cos(\omega_{i-3,i}) = \frac{a + bd_{i-3,i}^2}{c},$$

where $a, b, c \in \mathbb{R}$ and $d_{i-3,i} \in [\underline{d}_{r_{i-3}r_i}, \overline{d}_{r_{i-3}r_i}]$. The fact that a, b, c are precise numbers is a consequence of $\{r_{i-1}, r_i\}, \{r_{i-2}, r_i\} \in E'$.

Considering $\omega_{i-3,i} = 0$ and $\omega_{i-3,i} = 2\pi$, we get the minimum value for $\underline{d}_{r_{i-3}r_i}$, denoted by $d^{\min}_{r_{i-3}r_i}$, and the maximum value for $\overline{d}_{r_{i-3}r_i}$, denoted by $d^{\max}_{r_{i-3}r_i}$, respectively. Thus, $[\underline{d}_{r_{i-3}r_i}, \overline{d}_{r_{i-3}r_i}] \subset [d^{\min}_{r_{i-3}r_i}, d^{\max}_{r_{i-3}r_i}]$. When $d_{i-3,i}$ is a precise number ($d_{i-3,i} \in \mathbb{R}$), with $d^{\min}_{r_{i-3}r_i} < d_{i-3,i} < d^{\max}_{r_{i-3}r_i}$, we obtain two possible values for $\omega_{i-3,i}$, associated to two positions in $\mathbb{R}^3$ for r_i. However, when $d_{i-3,i} = [\underline{d}_{r_{i-3}r_i}, \overline{d}_{r_{i-3}r_i}]$, with $d^{\min}_{r_{i-3}r_i} < \underline{d}_{r_{i-3}r_i} < \overline{d}_{r_{i-3}r_i} < d^{\max}_{r_{i-3}r_i}$, we have now two possible intervals for $\omega_{i-3,i}$, associated to two arcs in $\mathbb{R}^3$ for r_i. In Figure 3, we illustrate these two arcs given as the intersection of two spheres (centered at x_{i-1}, x_{i-2} with radius $d_{i-1,i}, d_{i-2,i}$, respectively) and a spherical shell, defined by other two spheres with the same center x_{i-3} but with radius given by $\underline{d}_{r_{i-3}r_i}$ and $\overline{d}_{r_{i-3}r_i}$. This is the geometrical interpretation of the branching phase of BP.

Thus, any re-order corresponds to a DMDGP order, where some of the pairs $\{r_i, r_j\}$, with $|i - j| \geq 3$, may not correspond to precise distances, but rather to intervals.

The concept of a re-order was an important step to apply all the properties of the DMDGP as a mathematical model for problems related to 3D protein structure determination using NMR data. In the same paper that appeared the first re-order for such kind of problems [41], an extension of the BP algorithm, called iBP, was developed. The basic idea is to sample values from the intervals $[\underline{d}_{r_{i-3}r_i}, \overline{d}_{r_{i-3}r_i}]$, implying that the search space will not be anymore a binary

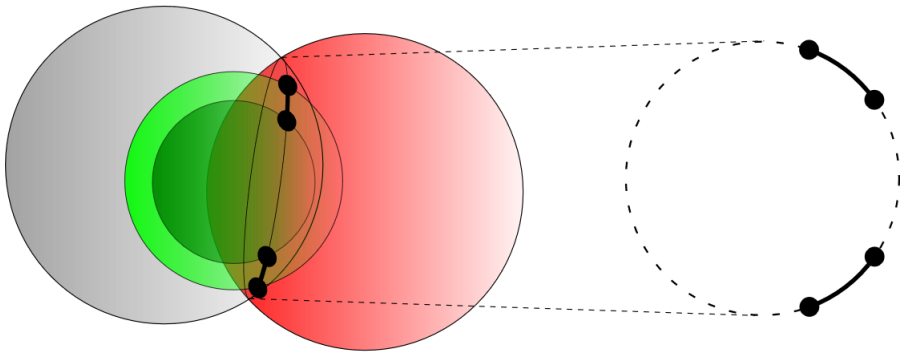

Figure 3: Geometric interpretation of branching in BP.

tree. Computational results presented in [12, 24] reveal the main difficult of the iBP: even for large samples, there is no guarantee that a solution will be found.

Essentially, there are two reasons for this difficulty:

1. The re-orders presented in [41, 24] have some pairs of vertices $\{r_{i-3}, r_i\}$ whose interval distances may not be associated to NMR data, implying no branches in the search tree, *i.e.*

$$[\underline{d}_{r_{i-3}r_i}, \overline{d}_{r_{i-3}r_i}] = [d^{\min}_{r_{i-3}r_i}, d^{\max}_{r_{i-3}r_i}]; \tag{3}$$

2. The sampling process "transforms" the iBP into a heuristic.

Very recent results [2, 3], using Clifford algebra, propose an alternative to avoid sampling process that allows iBP to explore the search space without "losing" solutions. However, in order to apply these new results to protein structure calculations, a new re-order must be defined to avoid the situation (3). The most important property of the re-order we will describe now is that it allows branches (in the iBP search) only at hydrogen atoms that are bonded to the protein backbone.

5.2. The Hand-Crafted (hc) Vertex Order

Let us define a protein graph $G = (V, E, d)$ associated to the backbone of a protein ($\{N^k, C^k_\alpha, C^k\}$, for $k = 1, ..., p$), including oxygen atoms O^k, bonded to

C^k, and hydrogen atoms H^k_α and H^k, bonded to C^k_α and C^k, respectively (see Figure 2, for $p = 3$).

The *hand-crafted vertex order* (*hc* order) we propose is the following:

$$\begin{aligned} hc \quad = \quad & \{N^1, H^1, H^{1'}, C^1_\alpha, N^1, H^1_\alpha, C^1, C^1_\alpha, ..., \\ & H^i, C^i_\alpha, O^{i-1}, N^i, H^i, C^i_\alpha, N^i, H^i_\alpha, C^i, C^i_\alpha, ..., \\ & H^p, C^p_\alpha, O^{p-1}, N^p, H^p, C^p_\alpha, N^p, H^p_\alpha, C^p, C^p_\alpha, O^p, C^p, O^{p'}\}, \end{aligned} \tag{4}$$

where $i = 2, ..., p-1$, $H^{1'}$ is the second hydrogen bonded to N^1 and $O^{p'}$ is the second oxygen bonded to C^p (Figure 4 illustrates this order for $p = 3$).

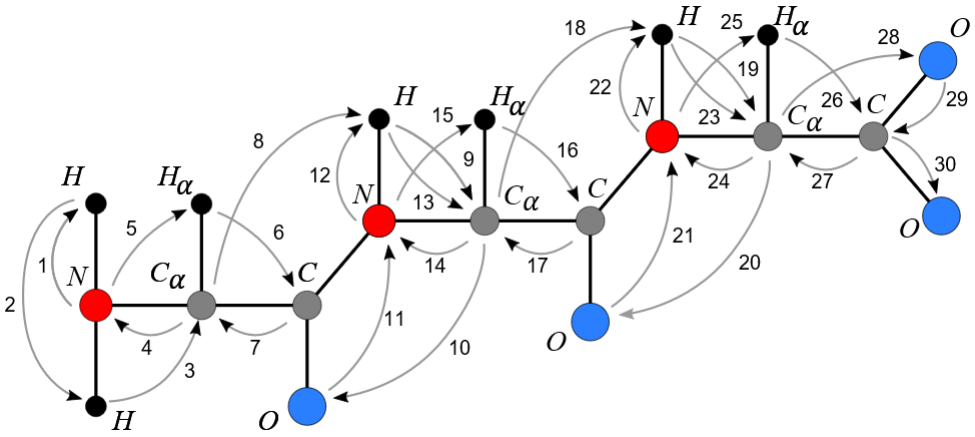

Figure 4: hc order.

We will prove now that *hc* is a re-order. We assigned the following order to the atoms of the first amino acid of a protein:

$$\{N^1, H^1, H^{1'}, C^1_\alpha, N^1, H^1_\alpha, C^1, C^1_\alpha\}. \tag{5}$$

Since we are assuming that all bond lengths and bond angles are fixed at their equilibrium values (the Rigid Geometry Hypothesis mentioned in the beginning of Section 5), the first and the second requirements of a re-order are satisfied. The third requirement is also satisfied, with the following distances

for $\{r_{i-3}, r_i\}$ (we will denote by $\mathbf{I}(\mathbf{H}^i, \mathbf{H}^j)$ the interval distance related to the pair of hydrogens $\{H^i, H^j\}$):

- $d(N^1, C^1_\alpha) \in (0, \infty)$,
- $d(H^1, N^1) \in (0, \infty)$,
- $d(H^{1'}, H^1_\alpha) = \mathbf{I}(\mathbf{H}^{1'}, \mathbf{H}^1_\alpha)$,
- $d(C^1_\alpha, C^1) \in (0, \infty)$,
- $d(N^1, C^1_\alpha) \in (0, \infty)$.

The nitrogen N^1 and the carbon C^1_α appear twice in the sequence, but they are related to the pairs $\{r_1, r_5\}$ and $\{r_4, r_8\}$.

To prove that *hc* is a re-order, we have to check the *connection* between the order (5) and the order for the second amino acid, given by the last three atoms of (5) and the six first atoms of the second amino acid:

$$\{H^1_\alpha, C^1, C^1_\alpha, H^2, C^2_\alpha, O^1, N^2, H^2, C^2_\alpha\}. \tag{6}$$

Here, in addition to the Rigid Geometry Hypothesis, we also have to use the properties of the so-called *Peptide Plane* [18], that says that the atoms $\{C^1_\alpha, C^1, O^1, N^2, H^2, C^2_\alpha\}$ are in the same plane (Figure 5). This implies that $d(C^1_\alpha, H^2)$ (related to the pair $\{r_8, r_9\}$), $d(C^1_\alpha, C^2_\alpha)$ (related to the pair $\{r_8, r_{10}\}$), $d(H^2, O^1)$ (related to the pair $\{r_9, r_{11}\}$), $d(C^2_\alpha, O^1)$ (related to the pair $\{r_{10}, r_{11}\}$), and $d(O^1, H^2)$ (related to the pair $\{r_{11}, r_{13}\}$) are all precise distances, satisfying the second requirement for a re-order. The third requirement is also satisfied, with the following distances for $\{r_{i-3}, r_i\}$:

- $d(H^1_\alpha, H^2) = \mathbf{I}(\mathbf{H}^1_\alpha, \mathbf{H}^2)$,
- $d(C^1, C^2_\alpha) \in (0, \infty)$,
- $d(C^1_\alpha, O^1) \in (0, \infty)$,
- $d(H^2, N^2) \in (0, \infty)$,

- $d(C_\alpha^2, H^2) \in (0, \infty)$,
- $d(O^1, C_\alpha^2) \in (0, \infty)$.

H^2 and C_α^2 are repeated, but they are related to the pairs $\{r_9, r_{13}\}$ and $\{r_{10}, r_{14}\}$, respectively.

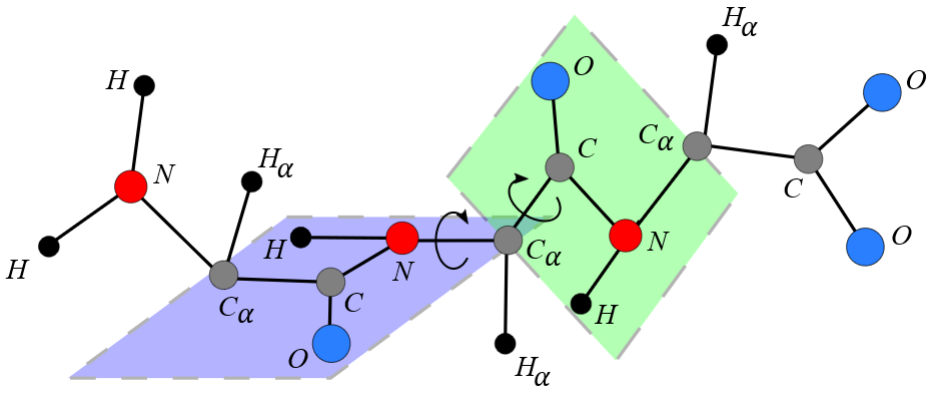

Figure 5: Peptide plane.

We assigned the following order to the atoms of a generic amino acid of a protein:

$$\{H^i, C_\alpha^i, O^{i-1}, N^i, H^i, C_\alpha^i, N^i, H_\alpha^i, C^i, C_\alpha^i\}. \tag{7}$$

By the same arguments used for the orders (5) and (6), the second and the third re-order requirements are satisfied, with the following distances for $\{r_{i-3}, r_i\}$:

- $d(H^i, N^i) \in (0, \infty)$,
- $d(C_\alpha^i, H^i) \in (0, \infty)$,
- $d(O^{i-1}, C_\alpha^i) \in (0, \infty)$,
- $d(N^i, N^i) = 0$,
- $d(H^i, H_\alpha^i) = \mathbf{I}(\mathbf{H^i}, \mathbf{H_\alpha^i})$,

- $d(C_\alpha^i, C^i) \in (0, \infty)$,
- $d(N^i, C_\alpha^i) \in (0, \infty)$.

In the order (7), H^i, C_α^i, and N^i are repeated, where H^i and C_α^i are related to pairs $\{r_i, r_j\}$, with $i - 3 < j$, and N^i is related to a pair $\{r_{i-3}, r_i\}$, which explains $d(N^i, N^i) = 0$ above.

The connection between two generic amino acids, given by

$$\{H_\alpha^i, C^i, C_\alpha^i, H^{i+1}, C_\alpha^{i+1}, O^i, N^{i+1}, H^{i+1}, C_\alpha^{i+1}\},$$

and the one between a generic amino acid and the last one, given by

$$\{H_\alpha^{p-1}, C^{p-1}, C_\alpha^{p-1}, H^p, C_\alpha^p, O^{p-1}, N^p, H^p, C_\alpha^p\},$$

have both the same order given in (6).

The result above implies the following distances for $\{r_{i-3}, r_i\}$, related to the connection between two generic amino acids,

- $d(H_\alpha^i, H^{i+1}) = \mathbf{I}(\mathbf{H}_\alpha^i, \mathbf{H}^{i+1})$,
- $d(C^i, C_\alpha^{i+1}) \in (0, \infty)$,
- $d(C_\alpha^i, O^i) \in (0, \infty)$,
- $d(H^{i+1}, N^{i+1}) \in (0, \infty)$,
- $d(C_\alpha^{i+1}, H^{i+1}) \in (0, \infty)$,
- $d(O^i, C_\alpha^{i+1}) \in (0, \infty)$,

and related to the connection between a generic amino acid and the last one:

- $d(H_\alpha^{p-1}, H^p) = \mathbf{I}(\mathbf{H}_\alpha^{p-1}, \mathbf{H}^p)$,
- $d(C^{p-1}, C_\alpha^p) \in (0, \infty)$,
- $d(C_\alpha^{p-1}, O^{p-1}) \in (0, \infty)$,
- $d(H^p, N^p) \in (0, \infty)$,

- $d(C_\alpha^p, H^p) \in (0, \infty)$,
- $d(O^{p-1}, C_\alpha^p) \in (0, \infty)$.

Finally, we assigned the following order to the atoms of the last amino acid of a protein:

$$\{H^p, C_\alpha^p, O^{p-1}, N^p, H^p, C_\alpha^p, N^p, H_\alpha^p, C^p, C_\alpha^p, O^p, C^p, O^{p'}\}. \tag{8}$$

Using once more the Rigid Geometry Hypothesis and the Peptide Plane Properties, the second and the third requirements of a re-order are satisfied, with the following distances related to $\{r_{i-3}, r_i\}$:

- $d(H^p, N^p) \in (0, \infty)$,
- $d(C_\alpha^p, H^p) \in (0, \infty)$,
- $d(O^{p-1}, C_\alpha^p) \in (0, \infty)$,
- $d(N^p, N^p) = 0$,
- $d(H^p, H_\alpha^p) = \mathbf{I}(\mathbf{H}_\alpha^p, \mathbf{H}^p)$,
- $d(C_\alpha^p, C^p) \in (0, \infty)$,
- $d(N^p, C_\alpha^p) \in (0, \infty)$,
- $d(H_\alpha^p, O^p) = \mathbf{I}(\mathbf{H}_\alpha^p, \mathbf{O}^p)$,
- $d(C^p, C^p) = 0$,
- $d(C_\alpha^p, O^{p'}) \in (0, \infty)$.

The distance $d(H_\alpha^p, O^p)$ is an interval, but the last level of the search tree can be related to the position of C^p, already determined using $d(C_\alpha^p, C^p)$.

The presented analysis can be summarized in the following theorem:

Theorem 1. *The hc order is a re-order.*

5.3. Minimal NMR Distance Information

In NMR experiments, the protein is submitted to a magnetic field, inducing the alignment of the magnetic moment of the observed nuclei. A resonance frequency characterizes the return of each perturbed magnetic moment, and the transmission of the perturbation is called Nuclear Overhauser Effect (NOE), which is proportional to d^{-6}, where d is the distance between two protons belonging to different atoms [13]. In general, if two protons are more than 5 Å apart, there is no NOE signal that can be measured for estimating their relative distance.

The measured signal recorded during NOE measurement may be distorted, due to dynamics of the sample protein, experimental noise, and the influence of neighboring atoms [52]. The NOE measurements are then converted into upper bounds and the corresponding lower bounds are given by the sum of the van der Waals radii of the involved atoms [33]. Thus, interval distances can be defined for hydrogen pairs that are close enough, implying the following result.

Theorem 2. *Using the hc order, the Rigid Geometry Hypothesis, and the Peptide Plane Properties, the set of distances between the pairs of hydrogen atoms*

$$\{H^{1\prime}, H^1_\alpha\}, ..., \{H^{i-1}_\alpha, H^i\}, \{H^i, H^i_\alpha\}, \{H^i_\alpha, H^{i+1}\}, ..., \{H^p, H^p_\alpha\}, \qquad (9)$$

where $i = 2, ..., p-1$ and p is the number of amino acids of a protein, are sufficient conditions to represent the solution space of the associated DGP as a search tree.

Let us consider this search tree more carefully. Since the *hc* order is a reorder, all distances $d_{i-1,i}$ and $d_{i-2,i}$ are precise values, greater than zero. Thus, concerning the size of the search space, we have to analyze all distances $d_{i-3,i}$ (remember that the branching of the search tree is the result of intersection between two spheres, with precise radius $d_{i-1,i}, d_{i-2,i}$, and another one with radius $d_{i-3,i}$, possibly given by an interval distance (Figure 3)).

In addition to the Rigid Geometry Hypothesis and the Peptide Plane Properties, we also need the *Chirality Property* [18], that defines the orientation of

the tetrahedrons formed by $\{N^1, H^1, H^{1'}, C_\alpha^1\}$ and $\{C_\alpha^i, N^i, H_\alpha^i, C^i\}$, implying only one possible position for C_α^1 and C^i, $i = 1, ..., p$ (Figure 6).

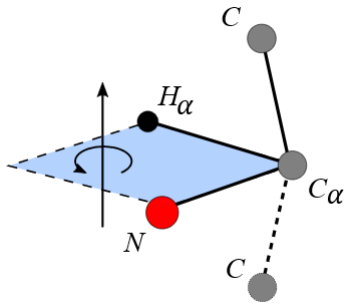

Figure 6: Chirality property.

Considering the first amino acid (with the links to the second one), we have:

- $d(N^1, C_\alpha^1) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for C_α^1, but **we can fix one** of them because of chirality defined on $\{N^1, H^1, H^{1'}, C_\alpha^1\}$.
- $d(H^1, N^1) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for N^1, but **we can fix one** of them, since N^1 is repeated.
- $d(H^{1'}, H_\alpha^1) = \mathbf{I}(\mathbf{H}^{1'}, \mathbf{H}_\alpha^1) \Rightarrow$ 2 possible arcs in $\mathbb{R}^3$ for H_α^1.
- $d(C_\alpha^1, C^1) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for C^1, but **we can fix one** of them because of chirality defined on $\{C_\alpha^1, N^1, H_\alpha^1, C^1\}$.
- $d(N^1, C_\alpha^1) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for C_α^1, but **we can fix one** of them, since C_α^1 is repeated.
- $d(H_\alpha^1, H^2) = \mathbf{I}(\mathbf{H}_\alpha^1, \mathbf{H}^2) \Rightarrow$ 2 possible arcs in $\mathbb{R}^3$ for H^2.
- $d(C^1, C_\alpha^2) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for C_α^2, but **we can fix one** of them because of the peptide plane already defined by $\{C^1, C_\alpha^1, H_\alpha^2\}$.
- $d(C_\alpha^1, O^1) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for O^1, but **we can fix one** of them because of the peptide plane already defined by $\{C^1, C_\alpha^1, H_\alpha^2\}$.

These are the distances $d_{i-3,i}$ in the generic amino acid (with the links to the next one):

- $d(H^i, N^i) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for N^i, but **we can fix one** of them because of the peptide plane already defined by $\{C^{i-1}, C_\alpha^{i-1}, H^i\}$.
- $d(C_\alpha^i, H^i) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for H^i, but **we can fix one** of them, since H^i is repeated.
- $d(O^{i-1}, C_\alpha^i) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for C_α^i, but **we can fix one** of them, since C_α^i is repeated.
- $d(N^i, N^i) = 0 \Rightarrow$ 1 possible position in $\mathbb{R}^3$ for N^i (the related torsion angle is 0).
- $d(H^i, H_\alpha^i) = \mathbf{I}(\mathbf{H}^i, \mathbf{H}_\alpha^i) \Rightarrow$ 2 possible arcs in $\mathbb{R}^3$ for H_α^i.
- $d(C_\alpha^i, C^i) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for C^i, but **we can fix one** of them because of chirality defined on $\{C_\alpha^i, N^i, H_\alpha^i, C^i\}$.
- $d(N^i, C_\alpha^i) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for C_α^i, but **we can fix one** of them, since C_α^i is repeated.
- $d(H_\alpha^i, H^{i+1}) = \mathbf{I}(\mathbf{H}_\alpha^i, \mathbf{H}^{i+1}) \Rightarrow$ 2 possible arcs in $\mathbb{R}^3$ for H^{i+1}.
- $d(C^i, C_\alpha^{i+1}) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for C_α^{i+1}, but **we can fix one** of them because of the peptide plane already defined by $\{C^i, C_\alpha^i, H_\alpha^{i+1}\}$.
- $d(C_\alpha^i, O^i) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for O^i, but **we can fix one** of them because of the peptide plane already defined by $\{C^i, C_\alpha^i, H_\alpha^{i+1}\}$.

Now, let us analyze the distances $d_{i-3,i}$ in the last amino acid (as we already mentioned, we are considering that the last level of the search tree is being related to the position of C^p):

- $d(H^p, N^p) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for N^p, but **we can fix one** of them because of the peptide plane already defined by $\{C^{p-1}, C_\alpha^{p-1}, H^p\}$.

- $d(C_\alpha^p, H^p) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for H^p, but **we can fix one** of them, since H^p is repeated.
- $d(O^{p-1}, C_\alpha^p) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for C_α^p, but **we can fix one** of them, since C_α^p is repeated.
- $d(N^p, N^p) = 0 \Rightarrow$ 1 possible position in $\mathbb{R}^3$ for N^p (the related torsion angle is 0).
- $d(H^p, H_\alpha^p) = \boldsymbol{I(H^p, H_\alpha^p)} \Rightarrow$ 2 possible arcs in $\mathbb{R}^3$ for H_α^p.
- $d(C_\alpha^p, C^p) > 0 \Rightarrow$ 2 possible positions in $\mathbb{R}^3$ for C^p, but **we can fix one** of them because of chirality defined on $\{C_\alpha^p, N^p, H_\alpha^p, C^p\}$.

The discussion above implies the following result.

Theorem 3. *Using the hc order, the Rigid Geometry Hypothesis, the Peptide Plane Properties, the Chirality Property, and the set of distances between the pairs of hydrogen atoms*

$$\{H^{1\prime}, H_\alpha^1\}, ..., \{H_\alpha^{i-1}, H^i\}, \{H^i, H_\alpha^i\}, \{H_\alpha^i, H^{i+1}\}, ..., \{H^p, H_\alpha^p\}, \tag{10}$$

where $i = 2, ..., p-1$ *and* p *is the number of amino acids of a protein, the branches in the search tree occur only at hydrogen atoms given by*

$$\{H_\alpha^1, ..., H^i, H_\alpha^i, ..., H^p, H_\alpha^p\}. \tag{11}$$

There are two main consequences of this theorem:

1. If the distances related to the pairs (10) are precise values, the search space of the associated DGP is finite, represented as a binary tree;
2. If the distances related to the pairs (10) are precise values and there is at least one additional distance (from NMR data) for each hydrogen in the list (11), there is only one DGP solution that can be found in linear time.

Although they are very strong hypothesis (precise and additional distances), this kind of information emphasizes its relationship and importance with the

cardinality of the DGP solution set and the computational complexity of the problem. Clearly, inaccuracy and lack of such information increase the size of the search space.

From the definition of the hc order (4) and from Theorem 6, we can note that the position of atom N^i depends on the position of atom H^i and that the position of atom C^i depends on the position of atom H^i_α. Since the protein backbone is determined by the torsion angles defined by $\{N^{i-1}, C^{i-1}_\alpha, C^{i-1}, N^i\}$ and $\{C^{i-1}, N^{i-1}, C^{i-1}_\alpha, C^i\}$ (the so-called (ϕ, ψ) angles), the term *minimal NMR distance information* is justified by the fact that we are requiring only NMR distances related to $d(H^i, H^i_\alpha)$ and $d(H^{i-1}_\alpha, H^i)$.

Since atoms H^i, H^i_α are in the same amino acid, the associated distance $d(H^i, H^i_\alpha)$ should not pose difficulty to be detected by NMR. Although atoms H^{i-1}_α, H^i are in consecutive amino acids, there is just one torsion angle (the one defined by $\{N^{i-1}, C^{i-1}_\alpha, C^{i-1}, N^i\}$) related to the position of H^i, because the Peptide Plane "forces" the torsion angle defined by $\{C^{i-1}_\alpha, C^{i-1}, N^{i-1}, C^i_\alpha\}$ to be π radians. In the worst case, supposing that the distance $d(H^{i-1}_\alpha, H^i)$ is not available, we can use the "implicit" information associated to the fact that the distance was not detected [1], *i.e.* $d(H^{i-1}_\alpha, H^i) \in [tol, d^{\max}]$, where $tol = 5$ Å (or other value related to the NMR precision) and $d^{\max}$ is the value associated to the torsion angle defined by $\{N^{i-1}, C^{i-1}_\alpha, C^{i-1}, N^i\}$ fixed to π radians.

6. New Research Directions

The contribution of this paper is related to how to combine information from protein geometry (Rigid Geometry Hypothesis, Peptide Plane, and Chirality) and NMR experiments in order to model the problem of 3D protein calculation using NMR data as a DMDGP that also considers interval distances.

From the results of this work, we select four new research directions:

1. Exploit the hc order to design pruning devices for the iBP;
2. Apply the hc order, with these developed pruning devices, to the Clifford algebra approach recently proposed in the literature;

3. Investigate the possibility to design new NMR experiments that focus on the accuracy of distances between hydrogen atoms used in the hc order;
4. Develop robust algorithms that can integrate all of the items above.

Regarding item 1, we have already some results as consequences of this paper, the information on lower and upper bounds to the backbone torsion angles provided by NMR chemical shifts [56], and information on hydrogen bonds defined between a hydrogen (the one bound to N) of one amino acid and the oxygen (bound to C) of another one:

- Since the position of atom O^{i-1} is determined by the position of atom H^i, hydrogen bond distances can be used to prune unfeasible positions of H^i;
- Since the position of atom N^i is also determined by the position of atom H^i, NMR chemical shift information on the torsion angle defined by $\{N^{i-1}, C_\alpha^{i-1}, C^{i-1}, N^i\}$ can be used to prune unfeasible positions of H^i;
- Since the position of atom C^i is determined by the position of atom H_α^i, NMR chemical shift information on the torsion angle defined by $\{C^{i-1}, N^{i-1}, C_\alpha^{i-1}, C^i\}$ can be used to prune unfeasible positions of H_α^i.

Of course, all the information related to the NMR distances

$$d(H^j, H^i), d(H_\alpha^{j-1}, H^i) \text{ and } d(H^{j-1}, H_\alpha^i), d(H_\alpha^j, H_\alpha^i),$$

where $j < i$, can also be used to prune unfeasible positions of H^i and H_α^i.

Acknowledgments

The authors would like to thank the Brazilian research agencies CNPq and FAPESP for their financial support. We are also thankful to Angela Gronenborn for discussions that clarify some ideas in the paper.

[1] A. Agra, R. Figueiredo, C. Lavor, N. Maculan, A. Pereira, and C. Requejo, International Transactions in Operational Research, 24 (2017) (2011), 1023–1040.

[2] R. Alves and C. Lavor, Geometric algebra to model uncertainties in the discretizable molecular distance geometry problem, Advances in Applied Clifford Algebra, 27 (2017), 439-452.

[3] R. Alves, C. Lavor, C. Souza, and M. Souza, Clifford algebra and discretizable distance geometry, Mathematical Methods in the Applied Sciences, (2017), 10.1002/mma.4422/.

[4] B. Anderson, P. Belhumeur, T. Eren, D. Goldenberg, S. Morse, W. Whiteley, and R. Yang, Graphical properties of easily localizable sensor networks, Wireless Networks, 15 (2009), 177–191.

[5] C. Bajaj, The algebraic degree of geometric optimization problems, Discrete and Computational Geometry, 3 (1988), 177-191.

[6] R. Benedetti and J.-J. Risler, Real Algebraic and Semi-algebraic Sets, Hermann, Paris, (1990).

[7] S. Billinge, P. Duxbury, D. Gonçalves, C. Lavor, and A. Mucherino, Assigned and unassigned distance geometry: applications to biological molecules and nanostructures, 4OR, 14 (2016), 337–376.

[8] L. Blumenthal, Theory and Applications of Distance Geometry, Oxford University Press, Oxford, (1953).

[9] H. Bodlaender, F. Fomin, A. Koster, D. Kratsch, and D. Thilikos, A note on exact algorithms for vertex ordering problems on graphs, Theory of Computing Systems, 50 (2012), 420–432.

[10] R. Carvalho, C. Lavor, and F. Protti, Extending the geometric build-up algorithm for the molecular distance geometry problem, Information Processing Letters, 108 (2008), 234–237.

[11] A. Cassioli, O. Gunluk, C. Lavor, and L. Liberti, Discretization vertex orders in distance geometry, Discrete Applied Mathematics, (2015) 197, 27–41.

[12] A. Cassioli, B. Bordiaux, G. Bouvier, A. Mucherino, R. Alves, L. Liberti, M. Nilges, C. Lavor, and T. Malliavin, An algorithm to enumerate all possible protein conformations verifying a set of distance constraints, BMC Bioinformatics, 16 (2015), 16–23.

[13] G. Clore and A. Gronenborn, Determination of three-dimensional structures of proteins and nucleic acids in solution by nuclear magnetic resonance spectroscopy, Critical Reviews Biochemistry Molecular Biology, 24 (1989), 479–564.

[14] R. Connelly, Generic global rigidity, Discrete and Computational Geometry, 33 (2005), 549–563.

[15] V. Costa, A. Mucherino, C. Lavor, A. Cassioli, L. Carvalho, and N. Maculan, Discretization orders for protein side chains, Journal of Global Optimization, 60 (2014), 333-349.

[16] G. Crippen and T. Havel, Distance Geometry and Molecular Conformation, Wiley, New York, (1988).

[17] A. Crum Brown, On the theory of isomeric compounds, Transactions of the Royal Society of Edinburgh, 23 (1864), 707–719.

[18] B. Donald, Algorithms in Structural Molecular Biology, MIT Press, Boston, (2011).

[19] Q. Dong and Z. Wu, A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances, Journal of Global Optimization, 22 (2002), 365–375.

[20] T. Eren, D. Goldenberg, W. Whiteley, Y. Yang, A. Morse, B. Anderson, and P. Belhumeur, Rigidity, computation, and randomization in network localization, IEEE Infocom Proc., 4 (2004), pp. 2673–2684.

[21] K. Gibson and H. Scheraga, Energy minimization of rigid-geometry polypeptides with exactly closed disulfide loops, Journal of Computational Chemistry, 18 (1997), 403–415.

[22] H. Gluck, Almost all simply connected closed surfaces are rigid, Lecture Notes in Mathematics, 438 (1975), 225–239.

[23] D. Gonçalves and A. Mucherino, Discretization orders and efficient computation of Cartesian coordinates for distance geometry, Optimization Letters 8 (2014), 2111-2125.

[24] D. Gonçalves, A. Mucherino, C. Lavor, and L. liberti, Recent advances on the interval distance geometry problem, Journal of Global Optimization, (2017), 10.1007/s10898-016-0493-6.

[25] J. Graver, B. Servatius, and H. Servatius, Combinatorial Rigidity, AMS, Providence, (1993).

[26] P. Güntert, Structure calculation of biological macromolecules from NMR data, Quarterly Reviews of Biophysics, 31 (1998), 145–237.

[27] T. Havel, I. Kuntz, and G. Crippen, The combinatorial distance geometry approach to the calculation of molecular conformation, Journal of Theoretical Biology, 104 (1983), 359–381.

[28] T. Havel and K. Wüthrich, A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of ^{1}H-^{1}H proximities in solution, Bulletin of Mathematical Biology, 46 (1984), 673-698.

[29] B. Hendrickson, Conditions for unique graph realizations, SIAM Journal on Computing, 21 (1992), 65–84.

[30] L. Henneberg, Statik der starren Systeme, Bergstræsser, Darmstadt, (1886).

[31] B. Jackson and T. Jordán, Connected rigidity matroids and unique realization of graphs, Journal of Combinatorial Theory Series B, 94 (2005), 1–29.

[32] B. Jackson and T. Jordán, On the rigidity of molecular graphs, Combinatorica, 28 (2008), 645–658.

[33] A. Kline, W. Braun, and K. Wüthrich, Studies by ^{1}H nuclear magnetic resonance and distance geometry of the solution conformation of the a-amylase inhibitor Tendamistat, Journal of Molecular Biology, 189 (1986), 377–382.

[34] G. Laman, On graphs and rigidity of plane skeletal structures, Journal of Engineering Mathematics, 4 (1970), 331–340.

[35] C. Lavor, R. Alves, W. Figueiredo, A. Petraglia, and N. Maculan, Clifford algebra and the discretizable molecular distance geometry problem, Advances in Applied Clifford Algebra, 25 (2015), 925-942.

[36] C. Lavor, L. Liberti, and N. Maculan, Computational experience with the molecular distance geometry problem, in Global Optimization: Scientific and Engineering Case Studies, J. Pintér, ed., Springer, Berlin, (2006), pp. 213–225.

[37] C. Lavor, A. Mucherino, L. Liberti, and N. Maculan, On the computation of protein backbones by using artificial backbones of hydrogens, Journal of Global Optimization, 50 (2011), 329–344.

[38] C. Lavor, J. Lee, A. Lee-St. John, L. Liberti, A. Mucherino, and M. Sviridenko, Discretization orders for distance geometry problems, Optimization Letters, 6 (2012), 783–796.

[39] C. Lavor, L. Liberti, N. Maculan, and A. Mucherino, The discretizable molecular distance geometry problem, Computational Optimization and Applications, 52 (2012), 115–146.

[40] C. Lavor, L. Liberti, N. Maculan, and A. Mucherino, Recent advances on the discretizable molecular distance geometry problem, European Journal of Operational Research, 219 (2012), 698-706.

[41] C. Lavor, L. Liberti, and A. Mucherino, The interval branch-and-prune algorithm for the discretizable molecular distance geometry problem with inexact distances, Journal of Global Optimization, 56 (2013), 855–871.

[42] L. Liberti, C. Lavor, and N. Maculan, A branch-and-prune algorithm for the molecular distance geometry problem, International Transactions in Operational Research, 15 (2008), 1–17.

[43] L. Liberti, C. Lavor, A. Mucherino, and N. Maculan, Molecular distance geometry methods: from continuous to discrete, International Transactions in Operational Research, 18 (2010), 33–51.

[44] L. Liberti, C. Lavor, J. Alencar, and G. Resende, Counting the number of solutions of ^{K}DMDGP instances, Lecture Notes in Computer Science, 8085 (2013), 224–230.

[45] L. Liberti, C. Lavor, N. Maculan, and A. Mucherino, Euclidean distance geometry and applications, SIAM Review, 56 (2014), 3–69.

[46] L. Liberti and C. Lavor, Six mathematical gems from the history of distance geometry, International Transactions in Operational Research, 23 (2016), 897–920.

[47] L. Liberti, B. Masson, J. Lee, C. Lavor, and A. Mucherino, On the number of realizations of certain Henneberg graphs arising in protein conformation, Discrete Applied Mathematics, 165 (2014), 213-232.

[48] A. Mucherino, C. Lavor, and L. Liberti, The discretizable distance geometry problem, Optimization Letters, 6 (2012), 1671–1686.

[49] A. Mucherino, C. Lavor, and L. Liberti, Exploiting symmetry properties of the discretizable molecular distance geometry problem, Journal of Bioinformatics and Computational Biology, 10 (2012), 1242009(1-15).

[50] A. Mucherino, C. Lavor, L. Liberti, and N. Maculan, eds., Distance Geometry: Theory, Methods, and Applications, Springer, New York, (2013).

[51] C. Mueller, B. Martin, and A. Lumsdaine, A comparison of vertex ordering algorithms for large graph visualization, IEEE Proc. of the 6th International Asia-Pacific Symposium on Visualization, (2007), pp. 141-148.

[52] M. Nilges, Calculation of protein structures with ambiguous distance restraints, Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities, Journal of Molecular Biology, 245 (1995), 645–660.

[53] S. Sallaume, S. Martins, L. Ochi, W. Gramacho, C. Lavor, and L. Liberti, A discrete search algorithm for finding the structure of protein backbones and side chains, International Journal of Bioinformatics Research and Applications, 9 (2013), 261–270.

[54] R. Santana, P. Larrañaga, and J. Lozano, Side chain placement using estimation of distribution algorithms, Artificial Intelligence in Medicine, 39 (2007), 49–63.

[55] J. Saxe, Embeddability of weighted graphs in k-space is strongly np-hard, Proc. of 17th Allerton Conference in Communications, Control and Computing, (1979), pp. 480–489.

[56] Y. Shen, F. Delaglio, G. Cornilescu, and A. Bax, TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts, Journal of Biomolecular NMR, 44 (2009), 213–223.

[57] M. Sitharam and Y. Zhou, A tractable, approximate, combinatorial 3D rigidity characterization, in the Fifth Workshop on Automated Deduction in Geometry, 2004.

[58] M. Souza, A. Xavier, C. Lavor, and N. Maculan, Hyperbolic smoothing and penalty techniques applied to molecular structure determination, Operations Research Letters, 39 (2011), 461-465.

[59] M. Souza, C. Lavor, A. Muritiba, and N. Maculan, Solving the molecular distance geometry problem with inaccurate distance data, BMC Bioinformatics, 14 (2013), S71–S76.

[60] J. Sylvester, Chemistry and algebra, Nature, 17 (1877), 284–284.

[61] T.-S. Tay and W. Whiteley, Generating isostatic frameworks, Structural Topology, 11 (1985), 20–69.