Universidade Federal de Santa Catarina

Centro de Blumenau
Departamento de Matemática

PIBIC
Relatório Final

Geometria de Distâncias e Álgebras Geométricas: novas perspectivas geométricas, computacionais e aplicações

Disposição de Robôs Móveis no espaço Euclidiano 3D: uma aplicação de Geometria de Distâncias

Guilherme Philippi (g.philippi@grad.ufsc.br),

Orientador: Felipe Delfini Caetano Fidalgo (felipe.fidalgo@ufsc.br).

8 de maio de 2020

Sumário

Abstract

In this paper, we study the Discretizable Molecular Distance Geometry Problem (DMDGP) applied to proteins, as well as the necessary tools for its comprehension, going from the graph theory to biomolecular structures. We, also, deal with some recent results on the ordering of a protein graph that composes the problem. The text concludes with a study of the algorithm described in the literature to solve the problem efficiently and a brief section of computer simulations.

Keywords: DMDGP, Distance geometry, Optimization.

Resumo

Neste trabalho, foram estudados o Discretizable Molecular Distance Geometry Problem (DMDGP) aplicado as proteínas, bem como as ferramentas necessárias para sua compreensão, passando da teoria de grafos às estruturas biomoleculares. Também lidamos com alguns resultados recentes sobre a ordenação do grafo da proteína que compõe o problema. O texto se encerra com um estudo sobre o algoritmo descrito na literatura para solucionar o problema de forma eficiente e uma brevê seção de simulações computacionais.

Palavras-chave: DMDGP, Geometria de Distâncias, Otimização.

1 Introdução

Existe uma relação muito forte entre a forma geométrica das moléculas orgânicas e suas funções em organismos vivos [1]. Outrora, em pesquisas sobre a molécula de DNA (ácido desoxirribonucleico), descobriu-se que essa era parte fundamental da produção de um dos pilares para a vida: a proteína. Esta é a estrutura básica que utilizamos para organizar nossas moléculas, gerando informação, ao possibilitarem um mecanismo funcional natural para a vida. Por exemplo, podemos citar o seu papel no transporte de oxigênio (hemoglobina), na proteção do corpo contra organismos patogênicos (imunoglobulina), com a catalização de reações químicas (apoenzima), além de outras inúmeras funções primordiais no nosso organismo [2].

Por conta dessa motivação tem-se esforços como o de Kurt Wüthrich, que propôs que utilizássemos experimentos de *Ressonância Magnética Nuclear* (RMN) para calcular a estrutura tridimensional de uma molécula de proteína (que lhe rendeu o premio Nobel da Química em 2002 [3]). Porém, a RMN não tem como resultado direto a estrutura tridimensional de uma proteína, mas sim distâncias entre átomos relativamente próximos que compõem a proteína — com inconvenientes erros associados, pois tratam-se de valores experimentais [4].

Para podermos calcular a estrutura de uma proteína a partir dessas distâncias, de forma estática, respeitando restrições de outras informações provenientes da física e química, surgira um novo problema na literatura conhecido como *Molecular Distance Geometry Problem* (MDGP), que é uma particularização do *Distance Geometry Problem* (DGP) [5]. Tal problema, munido de uma ordem conveniente para percorrer seus átomos (que garante uma discretização do espaço de buscas por soluções), pode ser discretizado, gerando o *Discretizable MDGP* (DMDGP).

Este último trata-se do nosso problema fundamental, que será melhor definido no Capítulo **??**. Para podermos compreendê-lo, introduzimos a teoria de grafos (no Capítulo 2), seguido das principais informações sobre as estruturas biomoleculares das proteínas (Capítulo **??**). Por último, apresentamos o principal algorítimo responsável pela solução do problema (Capítulo **??**), contendo algumas simulações computacionais.

A revisão bibliográfica completa pode ser encontrados no fim do documento, sendo devidamente citada durante o texto.

Materiais e Métodos

No que se segue, apresenta-se o estudo desenvolvido neste trabalho.

2 Teoria de Grafos

Esta seção tem como objetivo apresentar um breve resumo da *teoria de grafos*, tema amplamente estudado por diversos matemáticos e aplicado em diversas áreas do conhecimento como computação, engenharia e matemática [**?**].

2.1 Descoberta (Eureka!)

Costuma-se dizer que a teoria iniciou em 1736, com base no artigo publicado por Leonhard Euler sobre as 7 pontes de Königsberg [6] [**?**], representada na Figura 1. Conta a história que os moradores daquela região perguntavam-se sobre a possibilidade de atravessar todas as sete pontes do local sem ter que repetir alguma delas. Esse é um problema muito usado para introduzir o tema [**?**] — propõe-se o desafio de ligar todos os pontos de um desenho sem tirar o lápis do papel e sem passar duas vezes no mesmo ponto. Euler provou que isso não era possível ao formular matematicamente o problema, dando origem a esta teoria.

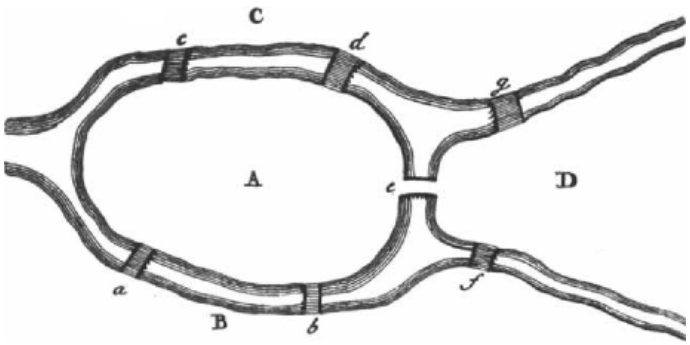

Figura 1: Ilustração original do problema [6].

A grande ideia de Euler foi abstrair o problema, vê-lo de uma forma elementar, como um conjunto de pontos conectados por curvas. Isso pode ser representado por um "gráfico", conforme a Figura 2 — é daí a origem do termo em inglês "Graph", que é tradução literal de "Gráfico". Essa representação facilita a análise e a busca por uma solução. Com isso, Euler percebeu que só seria possível solucionar o problema se houvessem exatamente nenhum ou apenas dois pontos conectados por um número impar de curvas (ou pontes) — o par de caminhos está associado com o ato de entrar e sair de um ponto [6]. Note que o caso de Köenigsberg, por tanto, não possui solução.

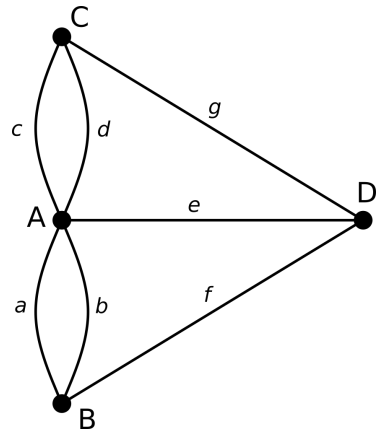

Figura 2: Grafo representando o caso da ponte de Köenigsberg.

Mas não podemos deixar todo o mérito com Euler. O conceito de grafo é muito intuitivo e foi proposto por diversas mentes brilhantes como formas de solucionar problemas que, em essência, são muito parecidos. Por exemplo, após Euler, a teoria foi redescoberta por Kirchhoff e Cayley [**?**]. Kirchhoff desenvolveu a teoria por volta de 1847, enquanto solucionava sistemas de equações lineares que relacionavam as correntes que percorriam as malhas de um circuito elétrico [**?**]. Dez anos depois, em 1857, foi a vez de Cayley que estudava diferentes estruturas em bioquímica formadas por carbonos (sempre com quatro ligações químicas) e hidrogênios (com apenas uma ligação), onde conseguiu formular seu problema introduzindo o conceito de *árvore* em grafos [**?**].

Além dessas, muitas outras situações reais podem ser convenientemente representadas por simples diagramas contendo um conjunto de pontos e um conjunto de relações entre pares desses pontos. Por exemplo, pode-se definir o conjunto $P = \{a, b, c\}$ das pessoas a, b e c e um conjunto $A = \{\{a, b\}, \{b, c\}\}$ como o conjunto de amizades entre essas pessoas — no caso, a é amigo de b, que é amigo de c, porém a não é amigo de c. Esta análise se torna muitíssimo útil quando se deseja estudar como uma informação se propaga em redes sociais.

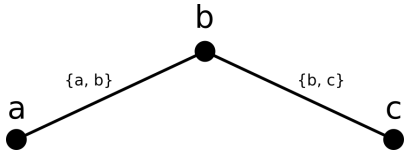

Figura 3: Grafo representando a relação entre as pessoas $\{a, b, c\}$.

2.2 Definição

Não há um forte consenso sobre as terminologias usadas pelos autores sobre grafos [**?**]. Essa confusão se deve tanto pela sua vasta disseminação em diversas áreas como pela enorme abstração que ela carrega. Cayley poderia chamar as relações entre pontos de ligações químicas enquanto Kirchhoff chamaria de curto-circuitos. No que se segue, há aqui um apanhado de definições sobre a teoria de grafos. Mas não sobre toda ela. Essa é uma grande área da matemática e não cabe aborda-lá

completamente nesse texto. Trata-se apenas do essencial para que o leitor possa progredir sem ter que consultar uma bibliografia complementar sobre grafos.

O ideal é começar pela definição geral.

> **Definição:** Um *grafo* G é uma tripla da forma (V_G, E_G, ψ_G), composta por um *conjunto de vértices* V_G, um *conjunto de arestas* E_G e uma *função de incidência* ψ_G que, por sua vez, associa a cada elemento de E_G um par não ordenado de elementos (nem sempre distintos) de V_G.

Nesse texto, porém, abstraiu-se a função de incidência ψ_G pois entende-se que o conjunto de arestas E_G é tal que, se $e \in E_G$, então $e = \{a, b\}$ onde $a, b \in V_G$. Fica implícita, por tanto, a associação dos elementos de V_G e E_G.

Aos elementos desses conjuntos (V_G e E_G), refere-se-os por *vértices* e *arestas*. Também, para um elemento $e \in E_G$, onde $e = \{u, v\}$, diz-se que u e v são *vértices adjacentes*. Chama-se u e e de *incidentes*, assim como v e e. À quantidade de vértices adjacentes a v da-se o nome *grau* de v. Para um vértice $v \in V_G$, define-se o *conjunto vizinhança* $N_G(v)$ como o conjunto de todos os vértices $u \in V_G$ adjacentes a v. Também, se duas arestas distintas e_1 e e_2 são incidentes com um vértice em comum, diz-se que e_1 e e_2 são *arestas adjacentes*.

Seja um grafo com m vértices e n arestas, dizer-se-há que este é um (m, n) *grafo*. Isto é, a Figura 3, para ilustrar, contém um $(3, 2)$ grafo onde os vértices a e b são adjacentes, assim como as arestas $\{a, b\}$ e $\{b, c\}$, porém, os vértices a e c não são. Define-se o $(1, 0)$ grafo como *trivial*.

Existem muitas variações de grafos [**?**]. Perceba que a definição de grafo permite *loops* (também chamado de *laço*, é uma aresta da forma $e = \{v, v\}$, ou seja, v é adjacente a si mesmo) e *múltiplas arestas* (mais do que uma aresta ligando os mesmos dois vértices). Grafos que não permitem múltiplas arestas ou loops são ditos *simples*. Grafos que permitem múltiplas arestas, mas não loops, são chamados de *multigrafos*. Caso também permitam os loops, os chamamos de *pseudografos*. Na Figura 2 (do problema das pontes de Köenigsberg) temos um multigrafo e na Figura 4 um pseudografo.

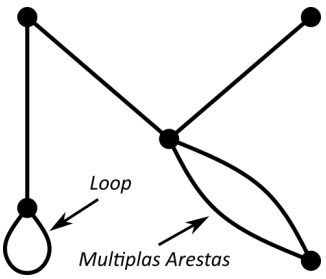

Figura 4: Exemplo de pseudografo contendo 5 vértices e 6 arestas.

Porém, para esse trabalho não interessa o estudo de multigrafos ou pseudografos. Por isso, adotou-se uma definição alternativa para grafos, visando restringir sua aplicação, como segue:

> **Definição:** Um *grafo simples* G é uma dupla da forma (V_G, E_G), composta por um conjunto não nulo e finito V_G e outro conjunto finito E_G de pares não ordenados de elementos **distintos** pertencentes a V_G.

2.3 Outras Definições Importantes

Diz-se que um grafo G é *rotulado* quando pode-se distinguir seus m vértices ao nomeá-los — algo como $v_1, v_2, \ldots, v_m$. Por exemplo, os grafos da Figura 5 são rotulados, enquanto o grafo da Figura 4 não é.

Dois grafos $G = (V_G, E_G)$ e $H = (V_H, E_H)$ são ditos *isomorfos* (escreve-se $G \cong H$) quando existe uma correspondência biunívoca entre os conjuntos de vértices V_G e V_H que preserve suas adjacências. A Figura 5 ilustra essa situação, com a correspondência $v_i \longleftrightarrow v_i$.

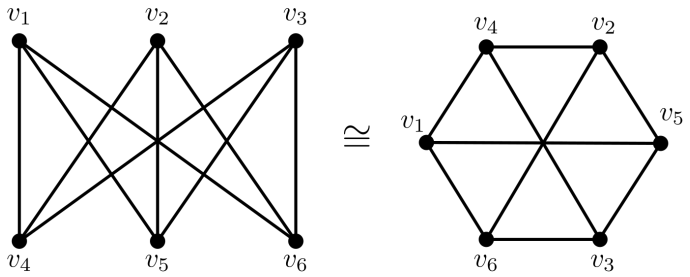

Figura 5: Diferentes representações isomórficas de um (6, 9) grafo.

O isomorfismo é uma relação de equivalência em grafos. Fica claro que, por mais que seja útil, a representação gráfica de um grafo existe apenas como um apelo didático. A geometria formada pelos vértices é escolha de quem desenha. Vários são os casos em que problemas envolvendo grafos são facilmente solucionáveis apenas rearranjando a forma como se desenha — como o caso das pontes de Köenigsberg. A resposta salta aos olhos.

2.3.1 Subgrafos

Diz-se que o grafo $G_1 = (V_{G_1}, E_{G_1})$ é *subgrafo* de $G = (V_G, E_G)$ se $V_{G_1} \subset V_G$ e $E_{G_1} \subset E_G$. Se G_1 é subgrafo de G, então G é *supergrafo* de G_1. Para qualquer $V \subset V_G$, existe um *subgrafo induzido* $\langle V \rangle$ definido por (V, E), onde $E \subset E_G$ contém todas as arestas $(v_1, v_2) \in E_G$ tal que $v_1, v_2 \in V$. Fica claro que dois vértices em $\langle V \rangle$ são adjacentes se, e somente se, forem também adjacentes em G.

Pode-se *remover um vértice* v de um grafo $G = (V_G, E_G)$, que resulta no subgrafo induzido $G - v = \langle V_G \setminus \{v\} \rangle$. Da mesma forma, pode-se *remover uma aresta* e de um grafo $G = (V_G, E_G)$, resultando no grafo $G - e = (V_G, E_G \setminus \{e\})$.

2.3.2 Caminhos

Um *passeio* em G é uma sequência finita não nula $W = v_0 e_1 v_1 e_2 v_2 \ldots e_k v_k$, onde seus termos são alternados entre vértices e arestas, tal que, para $1 \leq i \leq k$, antes e depois de e_i vem v_{i-1} e v_i, respectivamente. Diz-se que W é um passeio de v_0 para v_k, ou um (v_0, v_k)-passeio [?]. Os vértices v_0 e v_k são chamados origem e fim do passeio, respectivamente, e $v_1, v_2, \ldots, v_{k-1}$ são os vértices internos. O número k é o comprimento de W. Em um grafo simples, um passeio $v_0 e_1 v_1 e_2 v_2 \ldots e_k v_k$ é determinado suficientemente pela sequência dos vértices que o constitui $v_0 v_1 v_2 \ldots v_k$.

Se $W = v_0 v_1 \ldots v_k$ e $W' = v_k v_{k+1} \ldots v_l$ são passeios, o passeio $W^{-1} = v_k v_{k-1} \ldots v_0$ é dito *passeio reverso* de W e o passeio $WW' = v_0 v_1 \ldots v_l$ é dito *concatenação* de W

com W'. Chama-se *seção* do passeio W uma subsequência (v_i, v_j)-seção $= v_i v_{i+1} \dots v_j$ de termos consecutivos de W [?].

Se as arestas $e_1, e_2, \dots, e_k$ de um passeio W são todas distintas — o que sempre ocorre em grafos simples — chama-se W de *trilha*. Se, adicionalmente, os vértices da trilha W forem todos distintos, chama-se W de *caminho* (também conhecido como *caminho simples* [7]).

2.3.3 Conectividade

Dois vértices u e v de G são ditos *conectados* se existe um (u, v)-passeio em G. A conectividade induz uma relação de equivalência sobre o conjunto de vértices V [?]: Há uma partição de V em subconjuntos não vazios $V_1, V_2, \dots, V_\omega$ tal que dois vértices u e v são conectados se, e somente se, u e v pertencem ambos ao mesmo subconjunto V_i. Os subgrafos induzidos $\langle V_1 \rangle, \langle V_2 \rangle, \dots, \langle V_\omega \rangle$ são chamados *componentes de* G. Se G tem exatamente uma única componente, então G é dito *conectado*; e, do contrário, G é dito *desconectado*.

A Figura 6 mostra dois grafos: O grafo da esquerda é conectado — possui uma única componente $\langle \{v_1, v_2, v_3, v_4\} \rangle$; porém, o da direita não é — pois possui duas componentes $\langle \{v_1, v_2, v_3\} \rangle$, $\langle \{v_4\} \rangle$.

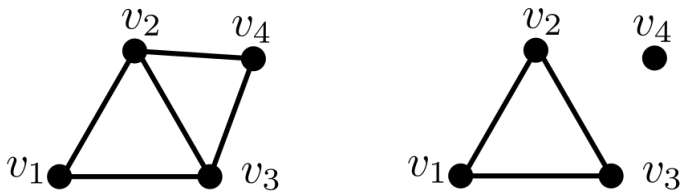

Figura 6: A esquerda um grafo conectado e, a direita, um grafo desconectado

2.3.4 Grafos Completos

Introduze-se agora uma classe especial de grafos [?]: Um grafo é dito *completo* se possui todas as suas arestas possíveis, i.e., para cada par de vértices distintos $u, v \in V_G$, u é adjacente a v (vide Figura 7).

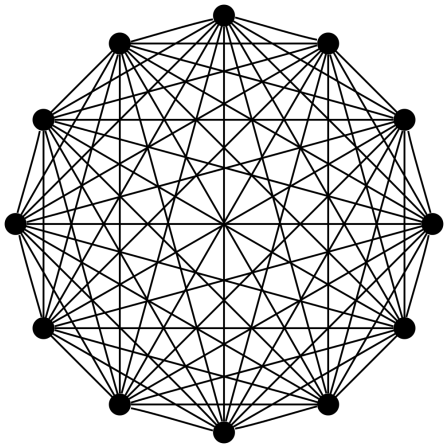

Figura 7: Diagrama de um grafo completo com 12 vértices ($|V| = 12$).

Em particular, chama-se de *k-clique* um grafo G com k vértices tal que G é completo. Por exemplo, se $k = 2$ tem-se uma linha — veja a Figura 8; e, se $k = 3$, forma-se um triangulo.

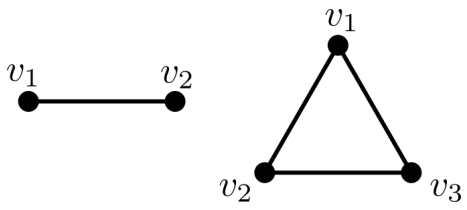

Figura 8: Em ordem: 2-clique e 3-clique.

2.4 Grafos Ponderados

As arestas $e \in E$ de um grafo G pode estar associadas com um número real $d(e)$, chamado de *peso da aresta* e [?]. Quando G tem todas as suas arestas associadas com pesos, define-se G como um *grafo ponderado*. Grafos ponderados são frequentemente associados com aplicações em teoria de grafos [7].

Costuma-se definir uma *função ponderação* $d : E \longrightarrow \mathbb{R}_+$ para mapear o conjunto de arestas E no conjunto dos números reais não negativos $\mathbb{R}_+$ [8]. Escreve-se $G = (V_G, E_G, d)$ como um grafo ponderado com o conjunto de vértices V_G, arestas E_G e função ponderação d.

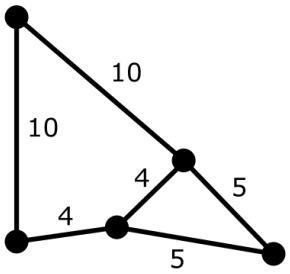

Figura 9: Representação de um grafo ponderado.

3 Geometria de Distâncias Euclidianas

Apresenta-se nesta seção uma introdução a *Geometria de Distâncias Euclidianas*. O nome "Geometria de Distâncias" diz respeito ao conceito desta geometria basear-se em distâncias ao invés de pontos. A palavra "Euclidiana" é importante para caracterizar as arestas — elementos fundamentais associados as distâncias — como segmentos, sem restringir seus ângulos de incidência [8].

3.1 Como tudo Começou

Por volta de 300 AC, Euclides de Alexandria organizou o conhecimento de sua época acerca da Geometria em uma obra composta por treze volumes, onde construiu, a partir de um pequeno conjunto de axiomas fortemente baseado nos conceitos de pontos e linhas, a chamada *Geometria Euclidiana* [9]. Em contraponto a visão original de Euclides, os primeiros conceitos geométricos usando *apenas distâncias* costumam estar associados aos trabalhos de Heron de Alexandria (10 a 80 DC) [8], com o desenvolvimento de um teorema que leva seu nome, como segue:

> ***Teorema de Heron*:** Sejam s o *semiperímetro* de um triângulo (se p é o perímetro, $s = \frac{p}{2}$) e a, b e c os comprimentos dos três lados deste triangulo. Então, a área A do triângulo é
>
> $$A = \sqrt{s(s-a)(s-b)(s-c)}. \qquad \text{(Fórmula de Heron)}$$

Pode-se dizer que esse foi o nascimento da *Geometria de Distâncias* (*Distance Geometry*, ou DG).

Algumas centenas de anos depois, em 1841, Arthur Cayley (1821 a 1895) generalizou a Fórmula de Heron através da construção de um determinante que calcula o conteúdo (volume n-dimensional) de um *simplex*[1] em qualquer dimensão [10]. Um século depois, em 1928, o matemático austríaco Karl Menger (1902 a 1985) re-organizou as ideias de Cayley e trabalhou em uma construção axiomática da geometria através de distâncias [11] — donde a alteração no nome do determinante de Cayley para como é conhecido hoje: "*Determinante de Cayley-Menger*".

> **Definição:** Sejam $A_0, A_1, \ldots, A_n$ $n+1$ pontos que definem os vértices de um n-simplex em um espaço euclidiano K-dimensional, onde $n \leq K$, e seja d_{ij} a distância entre os vértices A_i e A_j, onde $0 \leq i < j \leq n$. Então, o conteúdo v_n desse n-simplex é
>
> $$v_n^2 = \frac{(-1)^{n+1}}{(n!)^2 2^n} \begin{vmatrix} 0 & d_{01}^2 & d_{02}^2 & \ldots & d_{0n}^2 & 1 \\ d_{01}^2 & 0 & d_{12}^2 & \ldots & d_{1n}^2 & 1 \\ d_{02}^2 & d_{12}^2 & 0 & \ldots & d_{2n}^2 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ d_{0n}^2 & d_{1n}^2 & d_{2n}^2 & \ldots & 0 & 1 \\ 1 & 1 & 1 & \ldots & 1 & 0 \end{vmatrix}. \qquad \text{(Determinante de Cayley-Menger)}$$

Mas foi só com Leonard Blumenthal (1901 a 1984) que, em 1953, o termo Geometria de Distâncias foi cunhado — com a publicação de seu livro *"Theory and*

[1] Um simplex é uma generalização do conceito de triangulo a outras dimensões, i.e.: O *0-simples* é um ponto, *1-simplex* é um segmento de reta, *2-simplex* é um triangulo e o *3-simplex* é um tetraedro.

Applications of Distance Geometry" [12]. Blumenthal dedicou sua vida de trabalho para clarificar, organizar e traduzir as obras originais em alemão [8]. Ele acreditava que o problema mais importante nesta área era o *"Problema de Subconjunto"* (ou *Subset Problem*, originalmente), que consistia em encontrar condições necessárias e suficientes a fim de decidir quando uma matriz simétrica era, de fato, uma *matriz de distâncias*[2] [5]. Uma restrição desse problema à métrica euclidiana chama-se *Problema de Matrizes de Distâncias Euclidianas* (ou EDMP, do inglês *Euclidean Distance Matrix Problem*), como segue definida:

> **Problema de Matrizes de Distâncias Euclidianas:** Determinar se, para uma dada matriz quadrada $D_{n\times n} = (d_{ij})$, existe um inteiro K e um conjunto $\{p_1, \ldots, p_n\}$ de pontos em $\mathbb{R}^K$ tal que $d_{ij} = \|p_i - p_j\|$ para todo $i, j \leq n$.

Condições necessárias e suficientes para que uma matriz seja, de fato, uma matriz de distância euclidiana são dados em [13]. Para isso, apresenta-se um teorema onde se utiliza o Determinante de Cayley-Menger na criação de duas condições afirmando que, afim de $D_{n\times n}$ ser uma matriz de distâncias euclidianas, deve haver um K-simplex S de referência com conteúdo $v_K \neq 0$ em $\mathbb{R}^K$ e que todos os $(K+1)$-simplex e $(K+2)$-simplex contendo S como uma das faces devem estar contidos em $\mathbb{R}^K$ [5].

Blumenthal percebeu a importância em se respeitar as restrições métricas estabelecidas pelas matrizes de distâncias.

> *Quando temos como dado um conjunto de distâncias entre pares de pontos, a geometria das distâncias pode dar uma dica para encontrar um conjunto de coordenadas correto para pontos no espaço Euclideano tridimensional, satisfazendo as restrições de distâncias dadas.*
>
> (Blumenthal, 1953, [12])

Pode-se dizer que resolver o Problema de Matrizes de Distâncias Euclidianas está intimamente relacionado com descobrir as coordenadas dos pontos que definem suas distâncias. Perceba que este é um problema inverso, onde o "problema direto" correspondente é calcular distâncias associadas a pares de pontos dados. Note que este estudo tem enorme aplicabilidade [5].

Adiante, em 1979, Yemini (atualmente professor emérito de Ciência da Computação na Universidade de Columbia) foi o primeiro a flexibilizar a definição do EDMP ao considerar um conjunto de distâncias esparso [14, 5] — i.e., que não se tem todas as distâncias dadas a priori. Com isso, introduziu-se o que se chamou de *Problema Posição - Localização*, onde deseja-se calcular a localização de todos os objetos imersos em um espaço geográfico [14].

Assim, foi possível re-formular o problema fundamental de Geometria de Distâncias, o qual pode ser caracterizado de forma mais moderna pela utilização da Teoria de Grafos [5].

[2] Seja o par $(\mathcal{X}, d)$ um *espaço métrico* (vide Apêndice A), onde $\mathcal{X} = \{x_1, \ldots, x_n\}$. Uma *matriz de distância sobre* $\mathcal{X}$ é uma matriz quadrada $D_{n\times n} = (d_{uv})$ onde, para todo $u, v \leq n$, temos $d_{uv} = d(x_u, x_v)$ [5].

3.2 O Problema Fundamental

Uma *realização* é uma função que mapeia um conjunto de vértices de um grafo G para um espaço euclidiano de alguma dimensão dada [8].

> **Problema de Geometria de Distâncias (DGP):** Dados um grafo simples, ponderado e conectado $G = (V, E, d)$ e um inteiro $K > 0$, encontre uma realização $x : V \longrightarrow \mathbb{R}^K$ tal que:
>
> $$\forall \{u, v\} \in E, \quad \|x(u) - x(v)\| = d(u, v). \tag{1}$$

Desde que uma realização seja encontrada, também dá-se a ela o nome de *solução* do DGP. Por simplicidade — claramente um abuso de notação —, pode-se escrever x_u e d_{uv} no lugar de $x(u)$ e $d(u, v)$, respectivamente.

A principal diferença desta definição para o EDMP está acerca de que uma matriz de distância essencialmente representa um *grafo ponderado completo*. Em contraponto, o DGP não empoe qualquer estrutura em G[3], seguindo o conceito de matriz esparsa estabelecido por Yemini.

Por fim, na equação 1, utiliza-se a norma euclidiana $\| \cdot \|$ como métrica (ver Apêndice A), donde pode-se reescrever esta equação como

$$\forall \{u, v\} \in E, \quad \sqrt{\sum_{i=1}^{K} (x_{ui} - x_{vi})^2} = d_{uv}.$$

Como a definição de métrica garante a positividade das distâncias, pode-se esconder a raiz quadrada na equação acima, i.e.

$$\forall \{u, v\} \in E, \quad \sum_{i=1}^{K} (x_{ui} - x_{vi})^2 = d_{uv}^2. \tag{2}$$

3.3 Os Diferentes Problemas em DG

Em 2014, Leo Liberti *et al.* publicaram um ótimo compendio sobre a *Geometria de Distâncias Euclidianas e suas Aplicações* [5] e, em particular, desenvolveram um estudo taxonômico muito interessante sobre os problemas clássicos da área. No que se segue, devido a grande quantidade de siglas e variações dentro de DG, apresenta-se parte desse estudo, visando organizar os conceitos.

As principais aplicações em DG são no *calculo de estruturas moleculares* [15], na *localização de sensores em rede sem fio* (*Wireless Sensor Network Localization*, ou WSNL) [16], em *cinemática inversa* (*Inverse Kinematic*, ou IK) [17] e em *escalonamento multidimensional* (*Multidimensional Scaling*, ou MDS) [18].

3.3.1 Conformações Moleculares

Existe uma relação muito forte com a forma geométrica das moléculas e suas funções em organismos vivos [1]. Projetar drogas para curar uma doença específica se

[3] A menos, é claro, no que diz respeito a seus vértices estarem conectados. Porém, caso G não seja conectado, então ele consiste de um conjunto de diferentes subgrafos conectados, donde, a fim de solucionar o DGP, pode-se realizar cada subgrafo separadamente [8].

trata basicamente de conhecer o que uma certa proteína pode fazer em um organismo [8]. Proteínas se ligam em outras moléculas através do equilíbrio de forças agindo entre elas[4], por tanto, suas ligações dependem do seu formato.

Proteínas são constituídas por um grande conjunto de átomos e, alguns pares destes, trocam ligações químicas — sabe-se quais são esses átomos através de experimentos de cristalografia [20]. Então, se os átomos de uma molécula forem rotulados da forma $1, 3, 4, \ldots, n$, então é possível inferir:

- O conjunto de ligações $\{u, v\}$, onde u, v são átomos em $\{1, \ldots, n\}$;
- A distância entre u e v (para cara par ligado);
- O ângulo interno θ_v definido por duas ligações $\{u, v\}$ e $\{v, w\}$, com um átomo v em comum. (veja o Apêndice B)

Além desses dados, também é possível obter informações a partir de experimentos mais sofisticados, como a *Ressonância Magnética Nuclear* (RMN). Neste experimento é escolhida uma faixa de radiofrequência para bombardear uma amostra que está imersa em um campo magnético bastante intenso. Dependendo da radiofrequência utilizada (costuma-se usar a do hidrogênio), alguns núcleos atômicos irão absorver energia e outros não. Caso atinja-se uma frequência exata de ressonância dentro destes núcleos atômicos, é possível medir essa ressonância como um sinal de radiofrequência enviado dos núcleos atômicos — para calcular distâncias entre átomos próximos, com distâncias menores que 5Å.

De posse dessas informações, deseja-se realizar (localizar) todos os átomos da molécula. Esse problema, com todas as informações moleculares disponíveis, denomina-se *Estrutura Proteica a partir de Dados Brutos* (*Protein Structure from Raw Data*, ou PSRD)

Em particular, como as coordenadas atômicas pertencem ao $\mathbb{R}^3$, há uma particularização do DGP para o caso molecular, chamado *Problema de Geometria de Distâncias Moleculares* (*Molecular DGP*, ou MDGP). Trata-se do DGP com $K = 3$ fixo.

3.3.2 Localização de Sensores

O *Problema de Localização de Sensores em Rede sem Fio* (ou *WSNL Problem*) surge quando é necessário localizar um conjunto de objetos equipados com sensores eletrônicos capazes de medir distâncias entre si, geograficamente distribuídos, usando apenas medidas de distâncias entre pares destes objetos [16].

Por exemplo, *smartphones* com WIFI ativo podem criar uma rede conhecida por *Rede Ad-Hoc*, *i.e.*, eles conseguem criar uma rede para comunicar-se entre si, de forma *peer-to-peer*, sem a necessidade de uma torre central — cada aparelho funciona como uma pequena torre, de forma que a distância entre os aparelhos não pode ser excessiva. Dessa forma, os *smartphones* podem estimar a distância r de emparelhamento das suas conexões ao medir, por exemplo, qual a potência de

[4] Ou seja, o equilíbrio da energia potencial das moléculas, proporcional, principalmente, as variações nos comprimentos das ligações covalentes, as variações nos ângulos entre duas ligações covalentes consecutivas, as rotações sobre as ligações covalentes e as interações de van der Waals e interações eletrostáticas entre átomos [19].

transmissão do sinal, uma vez que sabe-se que a potência P de uma transmissão eletromagnética cai da forma

$$P = \frac{X}{r^n}, \tag{3}$$

onde X e n são constantes e dependem muito das condições do experimento, sendo obtidas experimentalmente [21].

Em essência, um problema do tipo WSNL segue a mesma definição do DGP, porém, com um subconjunto $A \subset V$ de vértices (chamados *âncoras*), onde os elementos de A tem uma posição em $\mathbb{R}^k$ dada a priori — isso é feito pois, normalmente, interessa saber a posição relativa de um objeto a outro, como é o caso do Sistema de Posicionamento Global, onde temos os satélites como âncoras e desejamos saber a posição dos aparelhos GPS em relação aos satélites.

Por motivos práticos — semelhantes ao caso molecular — as variações de interesse desse problema tem o K fixo em $K = 2$ ou $K = 3$. É comum, também, que se defina um WSNL como *solucionável* somente se seu grafo possua uma única realização válida [8] — noção conhecida como *globalmente rígido*: Diz-se que um grafo é *globalmente rígido* quando ele possui uma realização genérica x e, para todas as outras realizações x', x é congruente a x'.

3.3.3 Dinâmicas em Cinemática Inversa

Muito utilizada em robótica e animação computadorizada, a cinemática inversa cerne sobre mecanismos e seus movimentos rígidos, onde restringe-se os movimentos de forma a preservar a geometria do sistema. Sem o auxilio computacional e matemático a manipulação de mecanismos com muitos graus de liberdade pode ser inviável: Imagine a manipulação manual de cem vértices em uma haste simulando o comportamento de um braço articulado em uma animação. Com o auxílio da DG, um animador pode apenas configurar a posição final de um pequeno grupo de vértices (como os da extremidade da aresta, por exemplo) e um algorítimo de cinemática inversa é capaz de verificar se aquela posição é ou não viável e, se viável, qual a realização de todo o conjunto de vértices em razão da posição configurada [17].

Visando tal restrição mecânica, define-se o *Problema de Cinemática Inversa* (*Inverse Kinematic Problem*, ou IKP) como uma variação do WSNL — logo, tem o objetivo de descobrir posições em relação a certos pontos previamente realizados — com uma restrição no grafo que define o problema: deve ser um caminho simples com seus vértices finais sempre sendo âncoras [5].

3.3.4 Escalonamento Multidimensional

O problema de *Escalonamento Multidimensional* (*Multidimensional Scaling*, ou MDS) é definido como [5]: Dado um conjunto X de vetores, encontre um conjunto Y de vetores com menor dimensão (com $|X| = |Y|$) tal que a distância entre cada i-ésimo e j-ésimo vetores de Y tenham, aproximadamente, a mesma distância que seus pares de vetores correspondentes em X.

Esse problema é muito aplicado na analise de dados em Big Data [8]. É um meio de facilitar a visualização do nível de similaridade entre casos individuais — que não necessariamente precisam ter uma conexão aparente — em um conjunto de dados. Pode-se usá-lo, por exemplo, para visualizar em uma escala bidimensional

($\mathbb{R}^2$) a evolução da locomoção de animais no espaço tridimensional utilizando dados de séries temporais (espaço em diferentes tempos, logo, dados em $\mathbb{R}^4$).

3.4 A Busca de uma Solução

A abordagem mais simples, pode-se pensar, para encontrar um conjunto de soluções que satisfação a equação 2 é resolver o sistema de equações diretamente [22]. Infelizmente, para $K \geq 2$, há evidencias de que uma solução de forma fechada onde todo componente de x é expresso por raízes, não é possível [8].

No entanto, pode-se re-formular o problema como um problema de otimização global [8], onde o objetivo é minimizar a soma dos erros[5] entre as distâncias dadas a priori e as calculadas. Para isso, pode-se considerar uma única expressão que englobe todos os n erros, da forma

$$f(x_1, \ldots, x_n) = \sum_{(i,j)\in E} \left(\|x_i - x_j\| - d_{ij}\right)^2. \tag{4}$$

Fica claro que encontrar uma solução para o DGP é equivalente a encontrar realizações $x_i \in \mathbb{R}^3$, $i = 1, \ldots, n$, se e somente se $f(x_1, \ldots, x_n) = 0$ [8]. Pela definição de métrica (vide Apêndice A), 0 é o menor valor possível para f, donde diz-se que deseja-se *minimizar a função* $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ [22]. Ou seja,

$$\min_{x_i \in \mathbb{R}^n} f(x_1, \ldots, x_n). \tag{5}$$

E, no caso da métrica euclidiana [8], temos

$$\min_{x_j \in \mathbb{R}^n} \sum_{(u,v)\in E} \left(\sum_{i=1}^{K} (x_{ui} - x_{vi})^2 - d_{uv}^2 \right)^2. \tag{6}$$

Por tanto, a equação 6 tem como objetivo a minimização de um polinômio de múltiplas variáveis de grau quatro.

Um dos desafios da Otimização Global é que muitos dos métodos existentes — em especial, os mais eficientes — não garantem que uma otimização *global* será encontrada [8]. Isso se dá pois, dependendo do comportamento da função, existem muitos ótimos locais e os métodos não conseguem diferenciá-los de um global.

Infelizmente, essa abordagem via otimização é custosa do ponto de vista computacional [8]. Saxe demonstrou em 1979 [23] que resolver um DGP para qualquer dimensão — i.e., para qualquer valor de K — tem a complexidade computacional da classe **NP**-Hard. Em outras palavras, isso significa que a quantidade de mínimos locais de um DGP cresce exponencialmente proporcional a $|V|$ [24].

3.4.1 A Quantidade de Soluções do Problema

Pode-se verificar que, dada uma solução de um DGP, obtêm-se facilmente uma quantidade infinita (não enumerável!) de outras realizações válidas, distintas, através de rotações e translações desta solução inicial [22] — infelizmente, essas costumam não serem soluções de interesse.

[5] Em otimização, vê-se a equação 1 de forma não exata: $\|x_u - x_v\| = d_{uv} + \varepsilon$, onde ε é chamado *erro*. Ou seja, para minimizar o erro, precisa-se minimizar a expressão $f(x_u, x_v) = \|x_u - x_v\| - d_{uv}$.

Contudo, se for desconsiderado essas soluções advindas de translações e rotações, a quantidade de soluções do DGP depende da estrutura geométrica do grafo que a define: podem não haver nenhuma realização; uma única realização; uma quantidade finita (não única) de realizações; ou, um número incontável de realizações [8]. Perceba que o conjunto solução de um DGP, curiosamente, somente não pode ser um número infinito enumerável — resultado obtido através da *Geometria Algébrica Real* [25].

Por tanto, supondo que o conjunto solução de um DGP seja não vazio, sabe-se que ele é não enumerável ou finito. Se for finito, além de aplicar os métodos de Otimização Global — já definidos como custosos coputacionalmente —, pode-se explorar outras abordagens, como a Otimização Combinatória [22].

3.5 Combinatória do DGP

Nesta seção, analisando o espaço de busca por uma solução, faz-se um estudo sobre as condições que garantem a finitude do conjunto solução do problema.

Em particular, para um DGP definido em um espaço de dimensão K, a classe de problemas mais simples de resolver são dos grafos $(K+1)$-cliques, isto é, dos grafos completos com $K+1$ vértices [8].

3.5.1 Realização de Grafos Completos

Em um espaço unidimensional, i.e., uma reta, tem-se $K+1=2$. O grafo completo em dois vértices é $K_2 = (\{v_1, v_2\}, \{\{v_1, v_2\}\})$, ou seja, o grafo é dois vértices com uma ares entre eles — uma linha (veja a Figura 10). No espaço bidimensional, o grafo completo é um triangulo e, no tridimensional, um tetraedro.

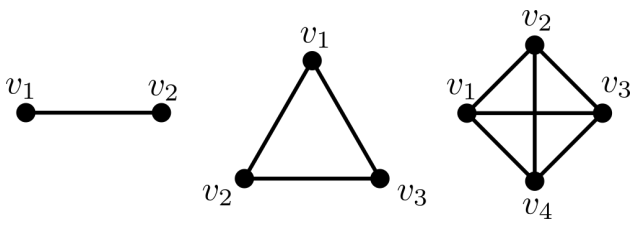

Figura 10: Em ordem: 2-clique; 3-clique; 4-clique.

No geral, se um $(K+1)$-clique tem uma realização em $\mathbb{R}^{K-1}$, ela é única a menos de rotações e translações [8]. É trivial, por tanto, que um $(K+2)$-clique tem, no máximo, uma realização no espaço $\mathbb{R}^K$ — caracterizado como solução do DGP. Por tanto, dependendo da geometria do grafo que define um DGP — isto é, se ele possuir cliques suficientes —, pode-se utilizar estas estruturas como "blocos básicos de construção" para planejar uma realização interativa do grafo como um todo [8].

3.5.2 Trilateração

Considere um 3-clique com $V = \{1, 2, 3\}$, onde $d_{12} = d_{23} = 1$ e $d_{13} = 2$. Então, uma possível realização sobre a linha real $\mathbb{R}$ que satisfaça todas as distâncias é $x_1 = 0$, $x_2 = 1$ e $x_3 = 2$. Uma forma de obter o valor de x_3 dado os valores de x_1

e x_2 e as distâncias d_{13} e d_{23} é a *trilateração*: sabendo que $d_{13} = \|x_3 - x_1\| = 2$ e $d_{23} = \|x_3 - x_2\| = 1$, tem-se

$$x_3^2 - 2x_1x_3 + x_1^2 = 4 \quad \text{e} \tag{7}$$

$$x_3^2 - 2x_2x_3 + x_2^2 = 1. \tag{8}$$

Subtraindo a equação 8 da 7, obtêm-se

$$2(x_1 - x_2)x_3 = x_1^2 - x_2^2 - 3 \;\Rightarrow\; 2x_3 = 4 \;\Rightarrow\; x_3 = 2.$$

E pode-se generalizar esse exemplo facilmente para $(K+1)$-cliques em $\mathbb{R}^{K-1}$ [8]: Precisa-se conhecer a posição de K vértices e as distâncias destes ao $(K+1)$-ésimo vértice, assim, pode-se realizar o $(K+1)$-ésimo vértice, em tempo linear, resolvendo um sistema de K equações como acima.

Isto é, sejam $x_1, \ldots, x_k \in \mathbb{R}^{K-1}$ as posições para K vértices de um $(K+1)$-clique e, para todo $j \leq K$, seja $d_{j,K+1}$ a distância associada com a aresta $\{j, K+1\}$. Seja $y \in \mathbb{R}^{K-1}$ a posição do $(K+1)$-ésimo vértice; então, y deve respeitar as K equações quadráticas $\forall j \leq K$, $\|y - x_j\|^2 = d_{j,K+1}^2$, com $K-1$ incógnitas $y_1, \ldots, y_{k-1}$:

$$\begin{cases} \|y\|^2 - 2x_1y + \|x_1\|^2 &= d_{1,K+1}^2 \\ & \vdots \\ \|y\|^2 - 2x_ky + \|x_k\|^2 &= d_{K,K+1}^2 \end{cases} \tag{9}$$

Para qualquer $h \leq K$. seja e_h a h-ésima equação no sistema de equações 9: pode-se tomas as diferenças e formar um novo sistema $\forall\, h < K$ $(e_h - e_K)$ contendo $K-1$ equações com $K-1$ incógnitas:

$$\begin{cases} 2(x_1 - x_K) \quad \cdot y &= \|x_1\|^2 - \|x_K\|^2 - d_{1,K+1}^2 + d_{K,K+1}^2 \\ & \vdots \\ 2(x_{K-1} - x_K) \cdot y &= \|x_{K-1}\|^2 - \|x_K\|^2 - d_{K-1,K+1}^2 + d_{K,K+1}^2 \end{cases} \tag{10}$$

Note que o sistema de equações 10 é um *sistema linear* da forma

$$Ay = b, \tag{11}$$

onde $A = (2a_{ij})$ é uma matriz quadrada $(K-1) \times (K-1)$ com $a_{ij} = x_{ij} - x_{Kj}$ para todo $i, j < K$, e $b = (b_1, \ldots, b_{K-1})^T$ com $b_i = \|x_i\|^2 - \|x_K\|^2 - d_{K,K+1}^2$ para todo $i < K$.

Diferentes métodos para solução de sistemas lineares como esse são encontrados na bibliografia [26, 8] — no geral, a escolha do melhor depende de propriedades da matriz A, como, por exemplo, quão esparsa ela é.

Em especial, se A não é uma matriz singular, então ela possui uma inversa A^{-1}. Logo, podemos obter a posição do $(K+1)$-ésimo vértice da forma

$$Ay = b \quad \Rightarrow \quad A^{-1}Ay = A^{-1}b \quad \Rightarrow \quad x_{K+1} = y = A^{-1}b. \tag{12}$$

Porém, se A é singular, isso quer dizer que as linhas $a_i = x_i - x_K$ (para $i < K$) não são todas linearmente independentes [26]. Essa situação mostra algumas propriedades geométricas interessantes [8]. Por exemplo, se $K = 2$, significa que

$x_1 - x_2 = 0 \Rightarrow x_1 = x_K$, ou seja, que o segmento entre x_1 e x_2 é um simples ponto. Se estamos imersos no $\mathbb{R}^{K-1} = \mathbb{R}$ (i.e., a reta real), geometricamente, a situação é que x_3 está posicionado ou a direita ou a esquerda de $x_1 = x_2$, mas não se pode escolher.

Também, se $K = 3$, a singularidade de A implica que o triangulo definido por x_1, x_2 e x_3 é um apenas um segmento no plano (caso o rank de A é 1) ou um simples ponto (caso o rank for 0). Nesse primeiro caso, x_4 pode estar posicionado em ambos os lados da linha que contém o segmento e, no segundo caso, x_4 pode estar em qualquer um dos pontos formados pela circunferência com centro $x_1 = x_2 = x_3$ e raio $d_{14} = d_{24} = d_{34}$. Essa característica geométrica vale para valores maiores de K [8]: a singularidade de A implica que há sempre múltiplas soluções para x_{K+1}.

Deve-se mencionar que, a partir da equação 9, podemos chegar no sistema linear 11, mas a volta não é verdadeira [8]. Em particular, se o sistema 9 tem uma solução, então o sistema 11 tem a mesma solução. Porém, mesmo que o sistema 9 não tenha solução, o sistema 11 sempre terá uma solução única — desde que A não seja singular. Por tanto, para verificar a factibilidade de uma solução x_{K+1} advinda do sistema linear 11, deve ser verificar se as distâncias aos K vértices foram respeitadas — ou seja, se $\|x_{K+1} - x_i\| = d_{i,(K+1)}$, para todo $i \leq K$.

3.5.3 Realização Iterativa de Grafos Completos

Uma característica interessante nos grafos completos é que, se (V, E) é um grafo completo, dado qualquer grafo induzido $\langle V' \rangle$ com $V' \subset V$, o subgrafo $\langle V' \rangle$ também é completo. Unindo esse princípio com a trilateração, nessa seção apresenta-se um algorítimo adaptado de [8] para a realização de grafos completos.

Primeiro, assume-se que exista um $(K+1)$-clique $K(G)$ em G, chamada clique inicial, que conhecemos a realização — em WSNL, por exemplo, comumente se utiliza nós ancoras como clique inicial [8]. Sem perda de generalidade, seja $\{1, \ldots, K+1\}$ o conjunto dos vértices os que formam a clique inicial, com realização $\{x_1, \ldots, x_{K+1}\}$. Seja, também, $N(i)$ o conjunto de vértices adjacentes ao i-ésimo vértice.

Algorithm 1: x = RealizacaoIterativa(G, d, K, x) [8]

```
   // Realize os próximos vértices iterativamente
 1 for i ∈ {K+2, ..., n} do
   |  /* Utilize a (K+1)-clique dos (K+1) antecessores imediatos
   |     de i para calcular a realização x_i.  Trilateracao() deve
   |     retornar ∅ caso o sistema não tenha solução            */
 2 |  x_i = Trilateracao(x_{i-K-1}, ..., x_{i-1});
   |  // verifique se x_i é factível com relação as outras
   |     distâncias
 3 |  for {j ∈ N(i) ; j < i} do
 4 |  |  if ||x_i - x_j|| ≠ d_ij then
   |  |  |  // Se não, marcar como não factível e sair do loop
 5 |  |  |  x_i = ∅;
 6 |  |  |  break;
 7 |  |  end
 8 |  end
 9 |  if x_i = ∅ then
   |  |  // Retornar que não foi possível concluir a realização
10 |  |  return ∅;
11 |  end
   |  // Retornar a realização factível
12 |  return x;
13 end
```

Uma informação muito importante sobre o Algorítimo 1 é que a complexidade de seu pior caso é $O(K^3 n)$ — para cada n vértices, deve-se resolver um sistema linear $K \times K$. Se não existir realização factível para G em $\mathbb{R}^K$, Algorítimo 1 retorna $\varnothing$.

Esse processo de trilateração em $\mathbb{R}^K$ é chamado K-*lateração* [27, 8].

4 Resultados e Discussão

Como vimos, a quantidade de soluções bem como o tempo para resolvê-las está crescendo de forma considerável proporcional a $|V|$. Isso é um problema real caso se queira calcular instâncias reais do problema, pois encontramos com grande facilidade proteínas com $|V|$ da ordem de 2000 no repositório wwPDB.

Uma boa proposta para implementações futuras seria um estudo sobre a otimização de memória necessária para implementar o BP. Isso pode ser observado no desenvolvimento do MD-Jeep [28], uma implementação em C feita por Antonio Mucherino, Leo Liberti e Carlile Lavor em 2010.

Uma solução trivial pensada para contornar esse aumento do número de soluções foi tentar manipular o valor de ε, de forma a produzir um filtro manual que diminuísse a quantidade de soluções. Porém, não obtivemos resultados satisfatórios. Pequenas oscilações em torno de um certo valor de ε (intrínsecos de cada molécula) faziam que, ou os resultados continuassem crescendo, ou não fossem nenhum.

Outra alternativa para solucionar esse problema pode ser encontrada estudando as simetrias do DMDGP [2] [5]. Perceba, nas Tabelas **??** e **??**, que os resultados possuem uma similaridade. Isso se dá devido as simetrias nas soluções de cada ramificação da arvore T, pois os resultados são simétricos (espelhados ao plano formado pelos três átomos anteriores [22]). Com isso, não precisamos buscar por todas as soluções da árvore de busca, pois, tendo uma solução, consegue-se a sua simétrica em tempo linear [2]. Isso é implementado em uma variação do BP, chamado *SymBP* [2].

Outras otimizações do algorítimo BP também podem ser encontrados na literatura, como uma versão que utiliza um paradigma Dividir e Conquistar [2], onde se constrói uma implementação paralela (Multithreading), que se utiliza das simetrias para produzir vários SymBP em paralelo. Um estudo sobre essas implementações poderiam ser úteis para otimizar nosso algorítimo.

5 Considerações Finais

Com isso concluímos um estudo elementar sobre o Discretizable Molecular Distance Geometry Problem, tal qual teve como resultado principal o software HCProt.

Nos cabe, nesse momento, voltarmos para as propostas levantadas internamente no início do projeto e verificar se elas foram cumpridas. Seguem o conjunto de objetivos específicos desse projeto, munidos de breve conclusão:

1. Entender as estruturas básicas de proteínas:

   Este fora feito de forma intensa, resultando no Capítulo **??** deste documento;

2. Relacionar-se eficientemente com o PDB (*Protein Data Bank*) - como extrair os dados computacionais que servirão de insumos:

   Apresentou-se este repositório junto do Capítulo **??**, devido sua proximidade temática. Vale mencionar que lá também fora apresentado o software HCProt, implementado pelo autor deste documento, que visa automatizar o processo de extração dos dados de distâncias do repositório PDB;

3. Compreender o DMDGP e sua estrutura de ordenamento dos vértices:

   Este objetivo se resume ao Capítulo **??**. Especialmente no estudo da ordenação HC, feita no fim do capítulo;

4. Conhecer todos os passos do algoritmo BP:

   Apresentado no Capítulo **??**, onde estudou-se todos os passos referentes a esse algorítimo: Inicialização, *branching* e *pruning*;

5. Simular, computacionalmente, o algoritmo BP com instâncias artificialmente geradas, como descrito na Literatura, dominando cada passo utilizado:

   Feito no fim do capítulo **??**, com resultados condizentes com os esperados;

6. Aplicar o Algoritmo BP estudado em estruturas proteicas como instâncias reais do problema:

   Não foi possível aplicar o BP a estás instâncias. Há uma pequena discussão, no Capitulo 4, sobre essa situação.

   Vale lembrar, porém, que uma parte importante desse resultado foi desenvolvido ao criar o software HCProt, possibilitando a extração dos dados de moléculas reais do repositório PDB, bem como a ordenação dos átomos, discretizando o problema.

Como implementações futuras, deseja-se estudar mais sobre as distâncias intervalares que aparecem com a ordenação HC [4] (detalhe este que não fora considerado na nossa implementação). Esses intervalos ocorrem justamente porque o HC order foi pensado de forma que a fase de *branching* do algorítimo BP só ocorra com distâncias extras advindas da RMN [4] — que, como vimos, são dados intervalares (Capítulo **??**). Uma ferramenta que tem demonstrado potencial para auxiliar nessa passagem é a *Geometria Conforme* [22].

Também deseja-se poder fazer um novo estudo do problema envolvendo a *Álgebra dos Quatérnios*, tentando otimizar o calculo que hoje envolve a matriz B_i, pois estes já demonstraram conter operações mais eficientes que as matriciais para realizar rotações [2].

Referências

[1] David L Nelson and Michael M Cox. *Lehninger principles of biochemistry*. W.H.Freeman and Company, 2013.

[2] Felipe Delfini Caetano Fidalgo. *Dividindo e conquistando com simetrias em geometria de distâncias*. PhD thesis, UNICAMP, Campinas, SP, Fevereiro 2015.

[3] K. Wüthrich. Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *American Association for the Advancement of Science*, 243(4887):45–50, Jan 1989.

[4] Carlile Lavor, Leo Liberti, Bruce Donald, Bradley Worley, Benjamin Bardiaux, Thérèse E Malliavin, and Michael Nilges. Minimal nmr distance information for rigidity of protein graphs. *Discrete Applied Mathematics*, 256:91–104, 2019.

[5] Leo Liberti, Carlile Lavor, Nelson Maculan, and Antonio Mucherino. Euclidean distance geometry and applications. *Society for Industrial and Applied Mathematics*, 56(1):3–69, February 2014.

[6] Leonhard Euler. Leonhard euler and the königsberg bridges. *Scientific American*, 189(1):66–72, 1953.

[7] Jayme Luiz Szwarcfiter. *Teoria computacional de grafos: Os algoritmos*. Elsevier Brasil, 2018.

[8] Leo Liberti and Carlile Lavor. *Euclidean Distance Geometry*. Springer, 2017.

[9] Irineu Bicudo et al. *Os elementos*. Unesp, 2009.

[10] Arthur Cayley. A theorem in the geometry of position. *Cambridge Mathematical Journal*, 2:267–271, 1841.

[11] Karl Menger. Untersuchungen über allgemeine metrik. *Mathematische Annalen*, 100(1):75–163, 1928.

[12] Leonard M Blumenthal. *Theory and applications of distance geometry*. Oxford University Press, Oxford, 1953.

[13] Manfred J Sippl and Harold A Scheraga. Cayley-menger coordinates. *Proceedings of the National Academy of Sciences*, 83(8):2283–2287, 1986.

[14] Yechiam Yemini. Some theoretical aspects of position-location problems. In *20th Annual Symposium on Foundations of Computer Science (sfcs 1979)*, pages 1–8. IEEE, 1979.

[15] Gordon M Crippen, Timothy F Havel, et al. *Distance geometry and molecular conformation*, volume 74. Research Studies Press Taunton, 1988.

[16] Yechiam Yemini. The positioning problem-a draft of an intermediate summary. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST, 1978.

[17] Deepak Tolani, Ambarish Goswami, and Norman I Badler. Real-time inverse kinematics techniques for anthropomorphic limbs. *Graphical models*, 62(5):353–388, 2000.

[18] Jan de Leeuw and Willem Heiser. 13 theory of multidimensional scaling. *Handbook of statistics*, 2:285–316, 1982.

[19] CC Lavor. *Uma abordagem determinística para minimização global da energia potencial de moléculas*. PhD thesis, PhD thesis, COPPE/UFRJ, Rio de Janeiro, 2001.

[20] GN Ramachandran, AS Kolaskar, C Ramakrishnan, and V Sasisekharan. The mean geometry of the peptide unit from crystal structure data. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 359(2):298–302, 1974.

[21] Andreas Savvides, Chih-Chieh Han, and Mani B Strivastava. Dynamic fine-grained localization in ad-hoc networks of sensors. In *Proceedings of the 7th annual international conference on Mobile computing and networking*, pages 166–179. ACM, 2001.

[22] C. Lavor, N. Maculan, M. Souza, and R. Alves. *Álgebra e Geometria no Cálculo de Estrutura Molecular*. IMPA, Rio de Janeiro, RJ, 31º colóquio brasileiro de matemática edition, 2017.

[23] James B Saxe. Embeddability of weighted graphs in k-space is strongly np-hard. In *Proc. of 17th Allerton Conference in Communications, Control and Computing, Monticello, IL*, pages 480–489, 1979.

[24] Carlile Lavor, Leo Liberti, Weldon A Lodwick, and Tiago Mendonça da Costa. *An Introduction to Distance Geometry applied to Molecular Geometry*. Springer, 2017.

[25] Riccardo Benedetti and Jean-Jacques Risler. In real algebraic and semi-algebraic sets. *Berlin, Hermann, Paris*, 1990.

[26] Elon Lages Lima. *Álgebra Linear*. SBM, Rio de Janeiro : IMPA, 1a edition, 2014.

[27] Tolga Eren, OK Goldenberg, Walter Whiteley, Yang Richard Yang, A Stephen Morse, Brian DO Anderson, and Peter N Belhumeur. Rigidity, computation, and randomization in network localization. In *IEEE INFOCOM 2004*, volume 4, pages 2673–2684. IEEE, 2004.

[28] Antonio Mucherino, Leo Liberti, and Carlile Lavor. Md-jeep: an implementation of a branch and prune algorithm for distance geometry problems. In *International Congress on Mathematical Software*, pages 186–197. Springer, 2010.

[29] Alfredo Steinbruch and Paulo Winterle. *Geometria Analítica*. Makron Books, São Paulo, SP, 2a edition, 1987.

A Métricas

Como esse texto utiliza fortemente o conceito de distância, é necessário e bem vindo que se gaste algum espaço para uma construção formal dessa ideia. A noção de distância está relacionada com o conceito de *métrica*, como segue.

Seja $\mathcal{X}$ um espaço vetorial K-dimensional sobre $\mathbb{R}$. *Métrica* é uma função de dois argumentos que mapeia pares ordenados de elementos em $\mathcal{X}$ para um número real não negativo. Precisamente, para todo x, y e $z \in \mathcal{X}$, uma função $d(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbf{R}$ é uma métrica se satisfaz os seguintes axiomas:

1. $d(x, y) = 0$ se, e somente se, $x = y$;
2. $d(x, y) = d(y, x)$;
3. $d(x, z) \leq d(x, y) + d(y, z)$;
4. $d(x, y) \geq 0$

Nesse trabalho, quando não é especificado qual métrica se está usando, fica implícita a utilização da *Métrica Euclidiana*, definida em função da *Norma Euclidiana*:

$$\forall x, y \in \mathcal{X}, d(x, y) = \|x - y\|_2 = \sqrt{\langle x, y \rangle} = \sqrt{\sum_{i=1}^{K} (x_i - y_i)^2}. \qquad \text{(Norma Euclidiana)}$$

O par $(\mathcal{X}, d)$ é chamado *espaço métrico*. A noção de métrica não depende de espaços vetoriais, donde pode ser facilmente generalizada fazendo $\mathcal{X}$ um conjunto qualquer.

B Lei dos Cos e Ângulos Entre dois Vetores no $\mathbb{R}^3$

A lei dos cossenos é uma propriedade trigonométrica válida para qualquer triângulo, permitindo encontrar o valor de um dos seus lados conhecendo apenas os outros lados e um ângulo. Porém, aqui utilizaremos a ideia reversa, onde, nesse caso, saberemos os lados e queremos descobrir os ângulos.

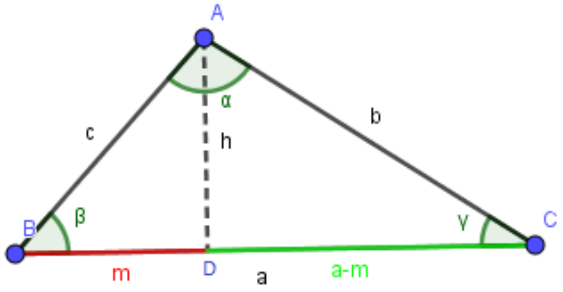

Figura 11: Triângulo para ilustrar a lei dos cossenos.

- **Demonstração Leis dos Cossenos:**

  Dado um triângulo qualquer, traça-se uma altura relativa ao lado a. Aplicando o *Teorema de Pitágoras* no ΔABD:

$$c^2 = m^2 + h^2 \rightarrow h^2 = c^2 - m^2 \tag{13}$$

  Aplicando novamente *Pitágoras*, porém, em ΔADC, obtemos:

$$b^2 = h^2 + (a - m)^2 \tag{14}$$

  Substituindo na equação 14 o valor de h^2 obtido em 13:

$$b^2 = c^2 - m^2 + a^2 - 2am + m^2$$

$$b^2 = c^2 + a^2 - 2am$$

  Analisando a Figura 11, pode-se perceber que $\frac{m}{c} = \cos\beta$, então:

$$b^2 = c^2 + a^2 - 2ac\cos\beta$$

  Analogamente, obtém-se:

$$c^2 = a^2 + b^2 - 2ab\cos\gamma$$

$$a^2 = b^2 + c^2 - 2bc\cos\alpha$$

  Note também que se o argumento dos cossenos for $\frac{\pi}{2}$ recaímos no Teorema de Pitágoras. ■

- **Ângulos Entre 2 Vetores:**

  Sejam dois vetores $\vec{u}$ e $\vec{v} \in \mathbb{R}^2$, representados na Figura 12

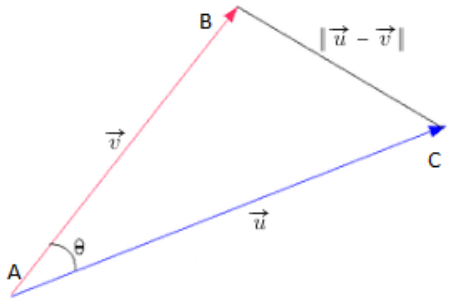

Figura 12: Diferença entre vetores u e v

  Para encontrarmos o angulo θ utilizaremos a lei dos cossenos aplicada a ΔABC:

$$\|\vec{u} - \vec{v}\|^2 = \|\vec{u}\|^2 + \|\vec{v}\|^2 - 2\|\vec{u}\|\|\vec{v}\|\cos\theta \tag{15}$$

  Utilizando a definição do produto escalar [29]

$$\|\vec{u} - \vec{v}\|^2 = \|\vec{u}\|^2 + \|\vec{v}\|^2 - 2\vec{u}\,\vec{v} \tag{16}$$

Comparando a equação 15 com a 16, obtemos trivialmente

$$\|\vec{u}\|^2 + \|\vec{v}\|^2 - 2\|\vec{u}\|\|\vec{v}\|\cos\theta = \|\vec{u}\|^2 + \|\vec{v}\|^2 - 2\vec{u}\,\vec{v}$$

$$\vec{u}\,\vec{v} = \|\vec{u}\|\|\vec{v}\|\cos\theta$$

Logo,

$$\cos\theta = \frac{\vec{u}\,\vec{v}}{\|\vec{u}\|\|\vec{v}\|}$$

■