Relatório final PIBIC - 2019/2020
Métodos numéricos para o problema de geometria de distâncias intervalar

Vinícius Douglas Cerutti
Orientador: Professor Douglas Gonçalves
Universidade Federal de Santa Catarina
Departamento de Matemática
Centro de Ciências Físicas e Matemáticas

Agosto de 2020

Resumo

O problema fundamental em Geometria de Distâncias consiste em determinar a posição de um conjunto de objetos utilizando apenas algumas distâncias entre pares de objetos. Este problema encontra importantes aplicações, como a determinação da estrutura de proteínas a partir de dados experimentais. Neste caso, as distâncias não são conhecidas exatamente, mas representadas por intervalos reais não-negativos, levando ao problema de Geometria de Distâncias Intervalar. Neste relatório, focamos no problema de otimização associado e empregamos o método do Gradiente Projetado Espectral para sua resolução. Como este é um método local, apresentamos uma relaxação convexa para o problema original, baseada em programação semidefinida, que permite a determinação de bons pontos iniciais para o método local. Experimentos numéricos mostram que a abordagem proposta é eficiente e capaz de recuperar, a partir de dados experimentais, estruturas proteicas com centenas de átomos em poucos minutos e com qualidade razoável.

Palavras-chave: Geometria de distâncias intervalar, matrizes de distâncias, estruturas moleculares, gradiente projetado espectral não monótono, programação semidefinida.

1 Introdução

Ao desenvolvermos modelos tri-dimensionais básicos de uma dada estrutura proteica podemos construir uma família de proteínas com características semelhantes, os quais, combinados com outras ferramentas de modelagem molecular, garantem um entendimento mais preciso da sua função e análise de seu comportamento enquanto interagem com demais proteínas [24]. Diversos estudos significativos estão sendo feitos nesse campo, bem como a criação de bancos de dados com diferentes estruturas proteicas, como por exemplo o "Protein Data Bank" [4], o que facilita o acesso a diferentes grupos de proteínas e permite uma melhor compreensão de suas semelhanças. É possível aplicar diferentes métodos experimentais para adquirir informações sobre a estrutura molecular de uma proteína. Até 1984, o método mais utilizado era a cristalografia e difração de raios-X [12].

Porém, com o trabalho de Kurt Wüthrich e seus co-pesquisadores iniciou-se uma revolução nesse campo, introduzindo o uso de ressonância magnética nuclear (RMN) [7], como um método conveniente para medir as distâncias interatômicas para proteínas em soluções aquosas, mais semelhantes aos ambientes naturais dos organismos vivos, do que os cristais usados na cristalografia por raios-X [28]. Entretanto, a constante variação entre tais medições resulta em modelos diferenciados de uma mesma estrutura, já que as distâncias obtidas no experimento de RMN não são exatas, mas representadas por intervalos de números reais não-negativos. A partir deste conjunto de distâncias intervalares procura-se então determinar as posições para os átomos da molécula, uma tarefa na qual as técnicas de Geometria de Distâncias desempenham um importante papel.

O problema fundamental de Geometria de Distâncias (PGD) é um problema inverso, em que uma lista de distâncias entre pares de objetos é utilizada para determinar a posição de cada um desses objetos em um espaço Euclidiano de dimensão K. Como o problema de Geometria de Distâncias também está associado ao problema de realizações de grafos, podemos então usar noções de teoria dos grafos para definir formalmente o problema fundamental de geometria de distâncias:

Definição 1.1 (PGD)**.** *Dado um inteiro positivo K e um grafo simples e não-orientado $\mathcal{G} = (V, E, d)$, cujos vértices estão ponderados por uma função não-negativa $d : E \to \mathbb{R}_+$, o Problema fundamental de Geometria de Distâncias consiste em determinar, se possível, uma função $x : V \to \mathbb{R}^K$ tal que*

$$\|x(u) - x(v)\| = d(\{u, v\}) \quad \forall \{u, v\} \in E, \tag{1.1}$$

em que $\|\cdot\|$ denota a norma Euclidiana.

Uma solução para o PGD é chamada de *realização* de $\mathcal{G}$. Sendo $n = |V|$ finito, observamos que o mapa x também pode ser representado como um vetor em $\mathbb{R}^{nK}$ ou como uma matriz $\mathbb{R}^{K\times n}$, aqui denominada por *matriz de realização*[1]. Resolver o problema é associar

[1]Em alguns momentos durante o texto o conceito de *matriz de realização* poderá ser referenciado como uma *realização* $X \in \mathbb{R}^{K\times n}$ omitindo-se o termo *matriz* do mesmo.

a cada vértice de $\mathcal{G}$ um único ponto em $\mathbb{R}^K$, satisfazendo as equações (1.1). Ou seja, ao posicionarmos os vértices u, v, temos que "acertar" a distância calculada $\|x(u) - x(v)\|$ com o valor dado $d(\{u, v\})$.

No entanto, como já discutido, métodos experimentais como RMN não são capazes de oferecer distâncias exatas $d_{u,v}$, mas sim, limitantes $[\underline{d}_{u,v}, \overline{d}_{u,v}]$, para *algumas* distâncias entre pares de átomos [16], i.e, nosso conjunto de distâncias nesse caso será incompleto e impreciso. O **Problema de Geometria de Distâncias molecular** (PGDm) surge nesse contexto. O PGDm consiste em determinar uma estrutura molecular de modo que sejam satisfeitas as restrições impostas sobre suas distâncias intermoleculares.

Podemos representar uma estrutura molecular por um grafo $\mathcal{G} = (V, E, d)$, ponderado e não-orientado, em que V representa o conjunto de átomos, E é uma relação simétrica em V a qual relaciona átomos cujas distâncias são conhecidas, e d representa as distâncias Euclidianas entre os átomos $\{u, v\} \in E$.

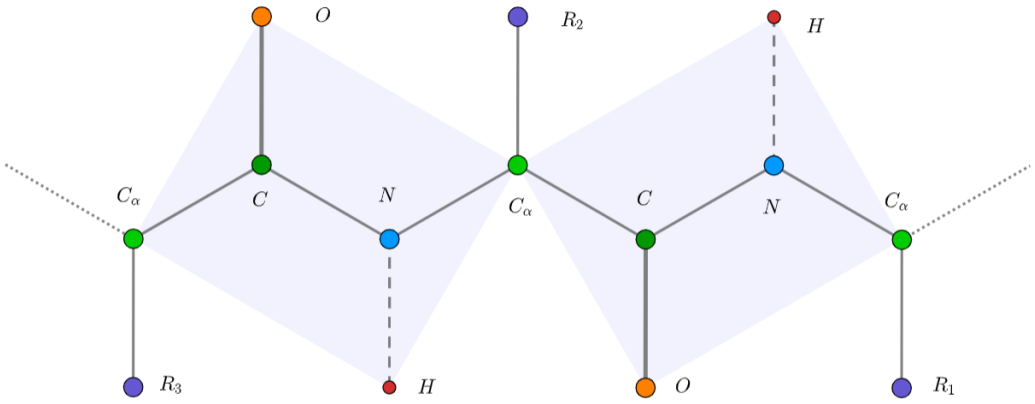


Figura 1.1: Um grafo $\mathcal{G}$ representado uma cadeia principal genérica de uma estrutura molecular.

Formalmente, o problema da geometria de distâncias molecular pode ser definido como um PGD em que visamos encontrar uma realização $x : V \to \mathbb{R}^K$, ou equivalentemente encontrar coordenadas cartesianas $x_u, x_v \in \mathbb{R}^K$ tais que

$$\underline{d}_{u,v}{}^2 \leq \|x_u - x_v\|^2 \leq \overline{d}_{u,v}{}^2, \ \forall \{u, v\} \in E. \tag{1.2}$$

Com base no artigo publicado por Glunt et al. [14], dados os limites superior $\overline{d}_{i,j}$ e inferior $\underline{d}_{i,j}$ para a distância entre os átomos i e j, o PGDm pode ser formulado como o seguinte problema de otimização [19]

$$\begin{aligned} \min_{X,y} . \quad & \sum_{i,j} (\|x_i - x_j\| - y_{i,j})^2 \\ \text{s.a} \quad & \underline{d}_{i,j} \leq y_{i,j} \leq \overline{d}_{i,j}, \ \forall \{i, j\} \in E. \end{aligned} \tag{1.3}$$

No primeiro ciclo PIBIC/CNPq 2018/2019, buscamos resolver o problema (1.3) utilizando o **Método do gradiente projetado espectral** (SPG[2]) [5]. Notamos uma performance muito boa do SPG quando iniciado em um ponto relativamente próximo da solução. Porém, na prática, raramente temos um ponto inicial próximo a um minimizador global. Assim, como o problema (1.3) é não-convexo, não temos a garantia de encontrar um minimizador global, e o sucesso do SPG depende bastante do ponto inicial. A fim de determinar um ponto inicial não arbitrário, que satisfaça (1.2) *aproximadamente*, neste trabalho consideraremos uma relaxação convexa para o PGDm baseada em programação semidefinida.

Este relatório está organizado da seguinte forma. Na Seção 2, trazemos uma breve revisão do método do gradiente projetado espectral. A Seção 3 apresenta a teoria sobre matrizes de distâncias Euclidianas e como estas se relacionam com matrizes positivas semidefinidas através do Teorema de Schoenberg. A formulação baseada em programação semidefinida é discutida na Seção 4 e experimentos numéricos em instâncias artificiais de proteínas são reportados na Seção 5. A Seção 6 traz as considerações finais.

2 O método do Gradiente Projetado Espectral

Seja $\Omega \subset \mathbb{R}^n$ um conjunto convexo, fechado e não vazio, e considere o problema de otimização:

$$\min_{x \in \Omega} . \quad f(x), \tag{2.1}$$

em que $f : \mathbb{R}^n \to \mathbb{R}$ é uma função continuamente diferenciável.

Por utilizar de pouca informação e pouco armazenamento em sua implementação, o gradiente projetado espectral (SPG) é adequado para problemas de grande porte do tipo (2.1). Discutimos abaixo as principais ideias deste método.

O método do Gradiente Espectral [5] une duas estruturas fundamentais: o *passo espectral* de Barzilai-Borwein (BB) [3] na direção contrária ao gradiente e a busca linear não-monótona de GLL [18]. Dentre as aplicações deste método, destacamos sua utilização na resolução do problema de otimização associado a determinação de estruturas moleculares [14].

O passo espectral λ_k relaciona o método espectral com a equação secante dos métodos *Quasi-Newton* para otimização irrestrita, os quais obedecem a seguinte relação recursiva:

$$x_{k+1} = x_k - \alpha_k H_k^{-1} \nabla f(x_k). \tag{2.2}$$

Agora, assuma que $H_{k+1} = \lambda_{k+1} I$ com $\lambda_{k+1} \in \mathbb{R}^*$ e considere $s_k = x_{k+1} - x_k$ e $g_k = \nabla f(x_{k+1}) - \nabla f(x_k)$. Assim, da equação secante [25] obtemos

$$\lambda_{k+1} s_k = g_k. \tag{2.3}$$

[2]Do inglês: Spectral Projected Gradient

Em geral, tal equação não admite solução. Sendo assim, tomando a solução de quadrados mínimos $\lambda_{k+1} = \text{argmin } \|\lambda s_k - g_k\|_2^2$, obtemos:

$$\lambda_{k+1} = \frac{\langle s_k, g_k \rangle}{\langle s_k, s_k \rangle} = \frac{s_k^\top g_k}{s_k^\top s_k}. \tag{2.4}$$

Logo, o método toma forma $x_{k+1} = x_k + \alpha_k d_k$ em que $d_k = -\lambda_k^{-1} \nabla f(x_k)$ e a expressão (2.4) é usada para gerar os coeficientes λ_k devidamente limitados, ou seja, o método usa salva-guardas $0 < \delta_{min} < \delta_{max} < \infty$ e define, a cada iteração

$$\lambda_k = \max\{\delta_{min}, \min\{\delta_{max}, \frac{s_k^\top g_k}{s_k^\top s_k}\}\}. \tag{2.5}$$

Além disso o SPG utiliza de uma busca linear não monótona que não impõe uma diminuição no valor de função a cada iteração, dando mais prioridade para as características do passo espectral. A essência por trás da estratégia é que, frequentemente, ao forçar monotonia nas iterações, podemos diminuir consideravelmente a taxa de convergência quando a iteração está presa perto de um vale estreito ou curvo da função [13], o que pode resultar em passos muito curtos ou um efeito de zig-zag na iteração. Por esse motivo, pode ser vantajoso permitir que a sequência dos iterados, ocasionalmente, gerem pontos com valores funcionais não monótonos [10].

Uma das primeiras iniciativas no assunto, foi desenvolvida por Grippo, Lamparielllo e Lucidi (GLL) [18] para métodos de Newton. Sua abordagem, fundamentalmente, foi a seguinte: dados *a priori* os parâmetros $\alpha_{min}, \alpha_{max}, \rho$ e δ com $0 < \lambda_1 < \lambda_2$ e $\rho, \gamma \in (0, 1)$, definimos $\alpha_k = \bar{\alpha_0}\rho^{h_k}$ em que $\bar{\alpha_0} \in [\alpha_{min}, \alpha_{max}]$ é o *passo inicial* (normalmente $\bar{\alpha_0} = 1$) e h_k é o menor inteiro não negativo tal que

$$f(x_k + \alpha_k d_k) \leq \max_{0 \leq j \leq m_k} f(x_{k-j}) + \gamma \alpha_k \langle \nabla f(x_k), d_k \rangle, \tag{2.6}$$

em que m_k é uma sequência de inteiros não-decrescente e limitada por um inteiro conhecido M, $m_0 = 0$ e para $k > 0$, $0 \leq m_k \leq \min\{m_{k-1} + 1, M\}$. Considerando as características desse método, suas propriedades de convergência [22], e seu bom desempenho em problemas de otimização irrestrita [21], surgiu o interesse em desenvolver uma variante do método espectral para problemas de otimização restrita, em particular, quando o conjunto viável Ω é convexo, fechado e não vazio. Assim surgiu o **Método do Gradiente Projetado Espectral** [5], descrito pela iteração

$$x_{k+1} = x_k + \alpha_k d_k,$$

em que a direção de descida d_k é tomada como

$$d_k = P_\Omega(x_k - \lambda_k \nabla f(x_k)) - x_k,$$

P_Ω denota a Projeção Euclidiana sobre Ω, λ_k é o passo espectral de BB (2.5) e α_k é um escalar satisfazendo a condição (2.6). Um pseudo-código do SPG é apresentado no Algoritmo 1. Para maiores detalhes, consulte o survey [5].

Algorithm 1: Gradiente Espectral Projetado

Entrada: $Z_0 \in \mathbb{R}^m$, $0 < \delta_{\min} < \delta_{\max}$, $\varepsilon, \mu, \bar{\alpha_0} > 0$, $\gamma, \rho \in (0,1)$, $1 \leq M, N \in \mathbb{Z}$, $\bar{\alpha_0} \in [\alpha_{\min}, \alpha_{\max}]$

Inicialização: Faça $k = 0$

Enquanto: $k \leq N$ faça:

Passo 1. Se: $|f(Z_k)| \leq \varepsilon$. **Retorne** Z_k;

Passo 2. Se: $k = 1$, defina $\lambda_k = 1$;

Do contrário: defina $s_k = Z_k - Z_{k-1}$ e $g_k = G(Z_k) - G(Z_{k-1})$ e calcule

$$\lambda_k = \min\left(\delta_{\max}, \max\left(\delta_{\min}, \frac{\langle s_k, g_k \rangle}{\langle s_k, s_k \rangle}\right)\right)$$

Passo 3. Calcule $Z_k^+ = Z_k - {\lambda_k}^{-1} G(Z_k)$ e $\bar{Z}_k = P_\Omega(Z_k^+)$, então defina $d_k = \bar{Z}_k - Z_k$;

Passo 4. Se: $|\langle G(Z_k), d_k \rangle| \leq \mu$ **Retorne** Z_k;

Passo 5.(*Backtracking*) Defina $\alpha = \bar{\alpha_0}$ e $f_{\max} = \max\{f(Z_{k-j}) \mid 0 \leq j \leq \min(k, M)\}$.

Se: $f(Z_k + \alpha d_{k+1}) < f_{max} + \gamma\alpha \langle G(Z_k), d_k \rangle$, então defina $\alpha_k \leftarrow \alpha$.

Do contrário: defina $\alpha = \bar{\alpha_0}\rho$ e retorne ao **Passo 5.**

Passo 6. Atualize $Z_{k+1} = Z_k + \alpha_k d_k$, $k \leftarrow k + 1$ e vá para o **Passo 1.**

Retorne: Z_k

3 Matrizes de Distâncias Euclidianas

As estruturas e propriedades relacionadas a grafos e suas realizações podem ser bem representadas usando matrizes. Por esta razão, podemos interpretar PGDs como problemas diretamente relacionados a certas matrizes conhecidas na literatura como *Matrizes de Distâncias* [11].

Definição 3.1 (Matriz de distâncias)**.** *Seja D uma matriz $n \times n$ simétrica. Diremos que D é uma matriz de distâncias, quando suas entradas $d_{i,j}$ forem não-negativas para todo $i \neq j$ e iguais a zero se $i = j$.*

Definição 3.2 (Matriz de Distâncias Euclidiana)**.** *Uma matriz de distâncias $D \in \mathbb{R}^{n \times n}$ é dita uma matriz de distâncias Euclidiana se existe um inteiro positivo K e vetores $x_1, x_2, \cdots, x_n \in \mathbb{R}^K, K < n$ tais que*

$$D_{i,j} = \|x_i - x_j\|^2, \ \forall i, j. \tag{3.1}$$

Além disso, diremos que o menor inteiro K para o qual verifica-se (3.1) é a dimensão de realização.

Note que a principal diferença entre o problema de determinar se D é uma Matriz de Distâncias Euclidianas (EDM)[3] e o Problema de Geometria de Distâncias é que no PGD a

[3]Do inglês: Euclidean Distance Matrix.

dimensão de interesse K é fixada, enquanto que para D ser uma EDM, basta que exista uma realização em alguma dimensão.

Veremos a seguir como uma matriz de distâncias Euclidiana se relaciona a uma possível realização $X \in \mathbb{R}^{K\times n}$.

Definição 3.3. *Dados, $x_1, \ldots, x_n \in \mathbb{R}^K$, designamos por matriz de Gram a matriz de produtos internos associada a esses vetores, ou seja*

$$G = \left[x_i^\top x_j\right]_{i,j} = \begin{bmatrix} x_1^\top x_1 & x_1^\top x_2 & \ldots & x_1^\top x_n \\ x_2^\top x_1 & x_2^\top x_2 & \ldots & x_2^\top x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_n^\top x_1 & x_n^\top x_2 & \ldots & x_n^\top x_n \end{bmatrix} = X^\top X. \tag{3.2}$$

Além disso, denotaremos também por: $\mathcal{S}_+^n$ conjunto das matrizes Simétricas Positivas Semidefinidas e $\mathbb{EDM}^n$ o conjunto das matrizes de distâncias Euclidianas de ordem n. Note que, tanto $\mathbb{EDM}^n$ quanto $\mathcal{S}_+^n$ são subconjuntos de $\mathcal{S}^n$ o conjunto das matrizes simétricas de ordem n.

Schoenberg [23] notou que ao explorarmos a relação entre norma Euclidiana e produto interno, torna-se possível associar as matrizes de distâncias Euclidianas com o conceito de matrizes de Gram.

De fato, dada uma matriz de distâncias $D = [{d_{i,j}}^2]$, seja $\{x_1, \cdots, x_n\}$ um conjunto de vetores satisfazendo (3.1) e $X \in \mathbb{R}^{K\times n}$ a matriz de realização associada. Da relação entre norma Euclidiana e produto interno:

$$D_{i,j} = \|x_i - x_j\|^2 = \|x_i\|^2 - 2x_i^\top x_j + \|x_j\|^2. \tag{3.3}$$

Assuma também que X é centralizada, ou seja, $X\mathbf{e} = 0$, em que $\mathbf{e}^\top = (1, 1, \cdots, 1, 1)$ é um vetor de uns com dimensão apropriada. Assim, podemos rescrever (3.3) como $D = \mathrm{edm}(X)$, em que

$$\mathrm{edm}(X) \stackrel{\text{def}}{=} \mathrm{diag}(X^\top X)\mathbf{e}^\top + \mathbf{e}\mathrm{diag}(X^\top X)^\top - 2X^\top X, \tag{3.4}$$

e $\mathrm{diag}(*)$ denota um vetor coluna cujos elementos pertencem a diagonal da matriz atribuída. Além disso, sabemos que *transformações rígidas* (isso inclui, rotações, translações, reflexões ou suas combinações) em um determinado conjunto de pontos não alteram suas respectivas distâncias. Com efeito, podemos facilmente deduzir esses fatos através da relação (3.4), chegando ao seguinte resultado cuja demonstração pode ser encontrada em [11].

Proposição 3.1 (Invariância por transformações rígidas)**.** *Seja $X \in \mathbb{R}^{K\times n}$ uma matriz de realização e S uma transformação rígida. Sendo X_S a ação de S sobre X, temos que $edm(X_S) = edm(X)$.*

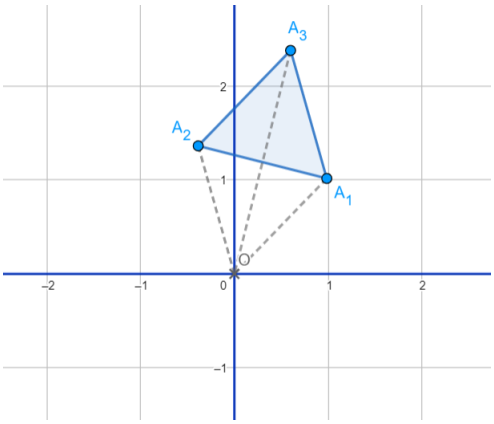


Figura 3.1: Um triângulo ($n = 3$) em $\mathbb{R}^2$ ($k = 2$). As linhas pontilhadas representam os vetores associados aos pontos A_1, A_2, A_3.

Perceba que, uma consequência dessa "invariância" é que, nem sempre é possível determinarmos a posição "absoluta" de um conjunto de pontos apenas baseado nas distâncias entre eles. De fato, consideremos o seguinte exemplo de uma EDM no caso em que $k = 2$:

$$D = [d_{i,j}{}^2] = \begin{bmatrix} 0 & d_{1,2}{}^2 & d_{1,3}{}^2 \\ d_{2,1}{}^2 & 0 & d_{2,3}{}^2 \\ d_{3,1}{}^2 & d_{3,2}{}^2 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 2 & 2 \\ 2 & 0 & 2 \\ 2 & 2 & 0 \end{bmatrix} \tag{3.5}$$

Nesse exemplo consideramos como estrutura original os vetores A_1, A_2 e A_3 presentes na Figura 3.1. Porém, note que mesmo possuindo todas as informações relacionadas as suas distâncias, tal estrutura não é única visto que qualquer rotação, reflexão ou translação desse conjunto de pontos igualmente produzirá a mesma edm(X), o que pode ser constatado na Figura 3.2.

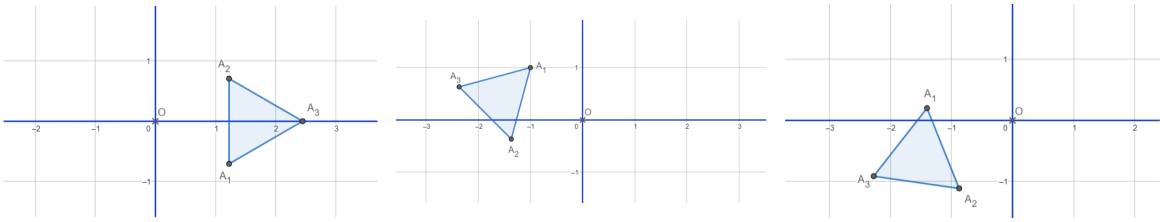


Figura 3.2: Transformações rígidas distintas que compartilham da mesma matriz de distâncias.

Seguindo de modo similar a construção de (3.4), podemos relacionar a EDM com a matriz de Gram associada $G = X^\top X$ através da seguinte relação

$$\mathcal{K}(G) \overset{\text{def}}{=} \text{diag}(G)\mathbf{e}^\top + \mathbf{e}\text{diag}(G)^\top - 2G. \tag{3.6}$$

Porém, note que a edm$(*)$ atua necessariamente sobre um conjunto/matriz de pontos $X \in \mathbb{R}^{K\times n}$ já concedido, enquanto que $\mathcal{K}$ atua sobre matrizes $G \in \mathcal{S}^n$, simétricas de ordem n,

sem a necessidade de um conhecimento prévio da localização de tais pontos. Note também que, $\mathcal{K}(G)$ é uma matriz de distâncias (não necessariamente Euclidiana) associada a matriz simétrica G. Pois bem, sabemos construir uma matriz de distâncias a partir de uma dada matriz $G \in \mathcal{S}^n$.

Veremos que o seguinte resultado nos oferece uma forma de obtermos G a partir de uma matriz de distâncias Euclidiana [1].

Proposição 3.2. *Sejam D uma EDM, G a matriz de Gram associada à D e seja J tal que $J = \left(I - \frac{1}{n}\boldsymbol{e}\boldsymbol{e}^\top\right)$. Então*

$$G = -\frac{1}{2}JDJ. \tag{3.7}$$

Através da Proposição 3.2 e as considerações anteriores, podemos caracterizar (3.6) como uma transformação linear, $\mathcal{K} : \mathcal{S}_c^n \to \mathbb{H}^n$ em que $\mathcal{S}_c^n$ representa o conjunto das matrizes simétricas de ordem n centralizadas na origem e $\mathbb{H}^n$ o conjunto das matrizes simétricas de ordem n com diagonal nula[4] [20]:

$$\mathcal{S}_c^n \stackrel{\text{def}}{=} \{Z \in \mathcal{S}^n \mid Z\mathbf{e} = 0\} \text{ e } \mathbb{H}^n \stackrel{\text{def}}{=} \{Z \in \mathcal{S}^n \mid \text{diag}(Z) = 0\},$$

e $\mathcal{K}^+ : \mathbb{H}^n \to \mathcal{S}_c^n$ dada por

$$\mathcal{K}^+(D) \stackrel{\text{def}}{=} -\frac{1}{2}JDJ, \tag{3.8}$$

representa a inversa generalizada de $\mathcal{K}$.

Schoenberg [23] e Young-Household [29] estabeleceram algumas propriedades básicas das matrizes de distâncias Euclidianas que foram precursoras para o desenvolvimento de futuras pesquisas em Geometria de Distâncias. Os estudos de Schoenberg possibilitaram a conexão das matrizes de Distâncias Euclidianas com a Programação Semidefinida (SDP[5]). O Teorema a seguir, demonstrado por Schoenberg [23], é fundamental nessa conexão.

Teorema 3.1 (Teorema de Schoenberg)**.** *Dada $D \in \mathbb{R}^{n\times n}$, D matriz de distâncias, D é EDM se, e somente se, $\mathcal{K}^+(D)$ é positiva semidefinida. Além disso, a dimensão de realização de D corresponde ao posto de $\mathcal{K}^+(D)$.*

Uma conclusão interessante do Teorema 3.1 bem como da equação (3.4) e sua discussão, é uma caracterização de EDM's [9] dada por

$$\mathbb{EDM}^n = \mathcal{K}(\mathcal{S}_+^n \cap \mathcal{S}_c^n).$$

[4]Do inglês: Hollow Matrices
[5]Do inglês: Semidefinite programming.

O Teorema 3.1 nos provém um método muito conveniente de obter uma realização $X \in \mathbb{R}^{K\times n}$ a partir de $\mathcal{K}^+(D)$, o que será útil para o desenvolvimento de algoritmos para o PGD.

Dada uma matriz de distância D, após calcular $G = \mathcal{K}^+(D)$ a partir de (3.8), a fim de determinarmos se D é uma EDM, é suficiente calcularmos a decomposição espectral $G = Q\Lambda Q^\top$, e verificar se todos os autovalores λ_i são não-negativos. Assumindo uma ordenação $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots \lambda_n = 0$, uma realização de D em $\mathbb{R}^r$ pode ser encontrada por

$$X = \sqrt{\Lambda(1:r, 1:r)}Q(:, 1:r)^\top, \tag{3.9}$$

em que $\Lambda(1:r, 1:r)$ é o primeiro bloco $r \times r$ de Λ e $Q(:, 1:r)$ denota a matriz formada pelas primeiras r colunas de Q. Além disso, se $G \succeq 0$, ou seja, é simétrica positiva semidefinida, porém a dimensão de realização considerada $K < r = \text{posto}(G)$, então, uma solução aproximada $\bar{G}$ pode ser obtida pela melhor aproximação de posto K [6] de G dada por $\bar{G} = \bar{X}^\top \bar{X}$ em que

$$\bar{X} = \sqrt{\Lambda(1:K, 1:K)}Q(:, 1:K)^\top. \tag{3.10}$$

Esta solução aproximada $\bar{X}$ estará na dimensão correta, mas possivelmente violará alguma das restrições de distâncias conhecidas.

4 Relaxação semidefinida

O Teorema 3.1 diz que se $\mathcal{K}^+(D) \in \mathcal{S}^n_+$, então o PGD associado a matriz de distâncias Euclidiana admite solução para todo $K \geq r = \text{posto}(\mathcal{K}^+(D))$. Entretanto, é importante salientarmos que para aplicar diretamente a Eq. (3.10), bem como as técnicas discutidas na Seção 3 em nosso problema original, é necessário conhecermos *todas* as entradas de D. Nesta seção, veremos como formular o PGD de (1.2) em termos de uma matriz simétrica G.

Suponha que garantimos a existência de uma matriz de Gram cuja matriz de realização esteja associada a EDM conhecida. Veja que podemos relacionar as distâncias contidas em nossa EDM com as entradas da matriz de Gram associada, pela expressão:

$$\begin{aligned} D_{i,j} = \text{edm}(X)_{i,j} = \|x_i - x_j\|^2 &= (x_i - x_j)^\top (x_i - x_j) \\ &= (Xe_i - Xe_j)^\top (Xe_i - Xe_j) \\ &= (e_i - e_j)^\top X^\top X (e_i - e_j), \end{aligned}$$

em que e_i denota o i-ésimo vetor canônico de $\mathbb{R}^n$. Ou seja, podemos relacionar o quadrado das distâncias $D_{i,j}$ diretamente com a matriz de Gram associada através da relação

$$D_{i,j} = (e_i - e_j)^\top G (e_i - e_j). \tag{4.1}$$

Além disso, considerando o produto interno entre duas matrizes $A, B \in \mathbb{R}^{m\times n}$ dado por

$$\langle A, B\rangle = \mathrm{Tr}(AB) = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{i,j}B_{i,j},$$

e as matrizes auxiliares $E_{i,j} = (e_i - e_j)(e_i - e_j)^\top$, podemos descrever o lado direito da igualdade (4.1) de uma forma ainda mais compacta utilizando da seguinte identidade:

$$(e_i - e_j)^\top G(e_i - e_j) = \mathrm{Tr}(GE_{i,j}).$$

Dados limitantes $[\underline{d}_{i,j}, \overline{d}_{i,j}]$ para a distâncias d_{ij}, essa identidade nos fornece uma maneira elegante de transferir tais restrições para a matriz de *Gram* associada. Obtemos uma reformulação equivalente para (1.2),

$$\underline{d}_{i,j}{}^2 \leq \langle G, E_{i,j}\rangle \leq \overline{d}_{i,j}{}^2, \ \forall\{i,j\} \in E. \tag{4.2}$$

Note que, na estrutura original de (1.2) todas as restrições impostas sobre a variável X são quadráticas, enquanto que em (4.2) as restrições são lineares em G. Portanto podemos garantir a existência de uma realização em $\mathbb{R}^K$ desde que consigamos uma matriz $G \in \mathcal{S}_c^n$ positiva semidefinida ($G \succeq 0$) que atenda as condições impostas, ou seja, uma possível solução para o seguinte problema de factibilidade:

$$\begin{aligned} \min_{G=G^\top} \quad & 0 \\ \text{S.a} \quad & \underline{d}_{i,j}{}^2 \leq \langle G, E_{i,j}\rangle \leq \overline{d}_{i,j}{}^2, \ \forall\{i,j\} \in E \\ & G \succeq 0, \ G\mathbf{e} = 0, \\ & \mathrm{posto}(G) = k. \end{aligned} \tag{4.3}$$

No entanto a restrição $\mathrm{posto}(G) = k$, torna o conjunto viável desse problema não-convexo, dificultando sua resolução. Assim, removendo essa restrição obtemos a seguinte *relaxação convexa* para o problema (4.3):

$$\begin{aligned} \min_{G=G^\top} \quad & \gamma\,\langle I, G\rangle \\ \text{S.a} \quad & \underline{d}_{i,j}{}^2 \leq \langle G, E_{i,j}\rangle \leq \overline{d}_{i,j}{}^2, \ \forall\{i,j\} \in E \\ & G \succeq 0, \ G\mathbf{e} = 0, \end{aligned} \tag{4.4}$$

em que o termo $\gamma\,\langle I, G\rangle$, com $\gamma < 0$ é adicionado a função objetivo como uma heurística [2] de redução de posto. Perceba que

$$\frac{1}{2n}\sum_{i,j}\|x_i - x_j\|^2 = \sum_i \langle x_i, x_i\rangle - \frac{1}{n}\sum_{i,j}\langle x_i, x_j\rangle \tag{4.5}$$

$$= \mathrm{Tr}(G) - \frac{1}{n}\langle G\mathbf{e}, \mathbf{e}\rangle = \mathrm{Tr}(G), \tag{4.6}$$

lembrando que $G\mathbf{e} = 0$. Logo ao maximizarmos a equação (4.5), buscamos afastar ao máximo os pontos x_i, aumentando o grau de dispersão destes pontos, na esperança que estes estejam próximos de uma variedade afim de dimensão pequena. [27].

A partir disso, podemos determinar um esquema geral que nos possibilitará encontrar soluções aproximadas para (1.2):

1. Se G^* é a solução ótima para o problema (4.5) com posto$(G^*) = r > K$, através de (3.9) obtemos uma realização que respeita todas as restrições impostas, porém, em um espaço r-dimensional.
2. Neste caso, uma solução aproximada $\bar{X}$ pode ser obtida pela melhor aproximação de posto k da expressão (3.10). Esta solução aproximada $\bar{X}$ estará na dimensão correta, entretanto, possivelmente terá algumas de suas restrições originais (1.2) violadas.
3. A fim de "satisfazer" as restrições perdidas, podemos aproveitar $\bar{X}$ para gerar um ponto inicial viável para o problema (1.3) com a finalidade de refinarmos a solução obtida previamente. Por exemplo, podemos escolher (X_0, y_0) em que $X_0 = \bar{X}$ e $y_0 = \bar{y}$ com

$$\bar{y}_{i,j} = \max\{\underline{d}_{i,j}, \min\{\|\bar{x}_i - \bar{x}_j\|, \overline{d}_{i,j}\}\}. \tag{4.7}$$

O esquema acima está organizado no Algoritmo 2, que retorna uma solução aproximada $\bar{X}$ para (1.2), com o vetor $\bar{y}$ correspondente.

Algorithm 2: Relaxação Convexa

Entrada: $D \in \mathbb{R}^{n\times n}$ matriz de distâncias e K dimensão de realização;
Passo 1. Resolva 4.4 obtendo G^* como solução;
Passo 2. Calcule a decomposição espectral de $G^* = Q\Lambda Q^\top$;
Passo 3
Se: $r = \text{posto}(G^*) \leq k$, considere a matriz de realização $\bar{X} = \sqrt{\Lambda(1:r, 1:r)}Q(:, 1:r)^\top$;
Do contrário: Determine a matriz de realização $\bar{X} \in \mathbb{R}^{K\times n}$ dada por

$$\bar{X} = \sqrt{\Lambda(1:K, 1:K)}Q(:, 1:K)^\top;$$

Passo 4. Calcule $\bar{y} \in \mathbb{R}^{|E|}$, em que

$$\bar{y}_{i,j} = \max\{\underline{d}_{i,j}, \min\{\|\bar{x}_i - \bar{x}_j\|, \overline{d}_{i,j}\}\},\ \forall\{i,j\} \in E;$$

Retorne: $(\bar{X}, \bar{y})$

A solução $\bar{X}$ pode ser posteriormente refinada, considerando a formulação (1.3). Para isso, utilizamos o par $(\bar{X}, \bar{y})$ como ponto inicial para um método local, por exemplo, o método de Gradiente Projetado Espectral (SPG) apresentado na Seção 2 (veja Algoritmo 1).

5 Experimentos Numéricos

Os experimentos computacionais para análise do método proposto na forma dos algoritmos de Relaxação Convexa (Algoritmo 2) e refinamento SPG (Algoritmo 1) foram realizados em instâncias artificiais, utilizando dos dados de proteínas presentes no banco de dados RCSB-Protein Data Bank [4].

5.1 Instâncias Artificiais

Para obtermos as distâncias entre os átomos realizamos um pré-processamento com base no primeiro modelo de cada proteína, e as instâncias artificiais foram geradas da seguinte maneira:

1. as distâncias entre os átomos de N, C_α e C de um mesmo aminoácido são mantidas exatas.
2. para pares de átomos que não se enquadram na característica 1, a distância entre eles é considerada apenas quando menor que 6Å. Neste caso, cria-se uma distância intervalar de tamanho Δ, tendo como centro a distância exata d:

$$\left[d - \frac{\Delta}{2}, d + \frac{\Delta}{2}\right].$$

Geramos então uma lista de pares de átomos E cujas distâncias obedecem as condições acima propostas.

Com o objetivo de avaliar as soluções apenas em respeito as restrições de distâncias, sem fornecer qualquer informação sobre a estrutura original X^*, usamos a avaliação do valor da função STRESS [14]:

$$\mathfrak{S}(X, y) = \sum_{\{i,j\} \in E} \left(\|x_i - x_j\| - y_{i,j}\right)^2, \text{ em que } \underline{d}_{i,j} \leq y_{i,j} \leq \overline{d}_{i,j}; \tag{5.1}$$

bem como o MDE (*Mean Distance Error*), o qual representa a média das violações das distâncias [17]:

$$MDE(X, [\underline{d}_{i,j}, \overline{d}_{i,j}]) = \frac{1}{|E|} \sum_{\{i,j\} \in E} \max\left(\frac{\underline{d}_{i,j} - d_{i,j}(X)}{\underline{d}_{i,j}}, 0\right) + \max\left(\frac{d_{i,j}(X) - \overline{d}_{i,j}}{\overline{d}_{i,j}}, 0\right), \tag{5.2}$$

Quando uma solução esperada X^* é conhecida, podemos também calcular o $RMSD$ (*Root Mean Square Deviation*) [17]:

$$RMSD(X, Z) = \min_{Q} \frac{1}{\sqrt{n}} \|QZ - X\|, \tag{5.3}$$

que representa o erro médio entre as posições ponto a ponto, de duas estruturas X e Z, centralizadas e alinhadas.

5.2 Detalhes de implementação

A implementação do Algoritmo 2 foi realizada com o software Matlab R2019a e utilizamos o solver SDPT3 [26] para solucionarmos o problema de programação semidefinida (4.4). O Método do Gradiente Projetado Espectral (SPG: Algoritmo 1) bem como o código geral que unifica os dois algoritmos foi implementado em Python 3.8. Todos os códigos podem ser acessados em [8].

Os testes foram realizados em um Notebook Dell, 16GB RAM, processador Intel core i7 2.2 GHz, 64 bit, sistema operacional Windows Home 10.

Na implementação do Algoritmo 1, utilizamos como parâmetros de salva-guarda $\delta_{min} = 10^{-16}, \delta_{max} = 10^{16}$, tolerâncias $\mu = \varepsilon = 10^{-6}$ e $\gamma = 10^{-4}$. Definimos como parâmetro da busca linear não-monótona (GLL) $M = 15$, valor de passo inicial $\bar{\alpha_0} = 1$, $\rho = 0.5$ e um número máximo de iterações $N = 2000$. Para a fase de relaxação convexa (Algoritmo 2) foram utilizados os parâmetros padrão do solver SDPT-3, os quais podem ser encontrados em [26].

5.3 Sensibilidade ao ponto inicial

Nesta seção estudaremos a *sensibilidade de escolha* do ponto inicial fornecido a métodos locais para o problema (1.3). Para isso, vamos considerar como ponto inicial $X^{(i)}$ uma perturbação da solução esperada X_{sol}, dada por

$$X^{(i)} = X_{sol} + \sigma N(0,1), \tag{5.4}$$

em que σ controla a amplitude do ruído e $N(0,1)$ é uma matriz $K \times n$, de modo que suas colunas são vetores aleatórios cujas entradas seguem uma distribuição normal padrão. A fim de estudar a influência do ponto inicial de (5.4) na resolução de (1.3), foram testadas algumas instâncias artificiais utilizando três amplitudes diferentes para σ, sendo estas $10^{-2}, 10$ e 10^2, todas as instâncias testadas foram criadas obedecendo os critérios estabelecidos em 5.1 utilizando como parâmetro $\Delta = 2/3$.

As Tabelas 1,2 apresentam os resultados dos testes para cada um dos três valores de σ apresentados. Nestas tabelas, PDB representa o indicador da proteína no banco de dados RCSB [4], N_a o número de átomos considerados da proteína listada, $|E|$ o número total de distâncias intervalares disponíveis e, para cada coluna com um valor de σ, apresentamos o valor da métrica avaliada no ponto inicial $X^{(i)}$ e na solução obtida pelo método SPG.

Note que, para $\sigma = 10$, mesmo havendo variações significativas na estrutura do ponto inicial considerado (como por exemplo, no caso das proteínas $2K35$ e $6HKC$), obtivemos uma aproximação de $RMSD$ na ordem de 10^{-1} com a solução esperada e a média da violação dos desvios das distâncias, MDE, permaneceu na ordem de 10^{-3}. Isso mostra que o método de SPG está realizando um bom trabalho em minimizar a violação das restrições, embora

			$\sigma = 10^{-2}$		$\sigma = 10$		$\sigma = 10^{2}$	
PDB	N_a	$\|E\|$	$RMSD_i$	$RMSD_f$	$RMSD_i$	$RMSD_f$	$RMSD_i$	$RMSD_f$
2Y2A	30	162	5.55e-03	5.56e-03	1.43e+01	4.42e+00	1.75e+02	4.79e+00
2JMY	103	1455	2.52e-03	2.46e-03	1.31e+01	2.81e+00	1.65e+02	5.18e+00
6G4U	164	2260	2.08e-03	2.09e-03	1.15e+01	1.75e+00	1.65e+02	6.38e+00
6HN9	225	3000	1.74e-03	1.69e-03	1.17e+01	2.11e+00	1.66e+02	5.57e+00
6FS5	272	3794	1.23e-03	1.07e-03	1.03e+01	2.77e+00	1.54e+02	7.22e+00
2K35	353	4085	5.87e-04	5.16e-04	1.00e+01	4.81e-01	1.60e+02	2.21e+00
1DT4	357	3488	1.23e-03	1.20e-03	9.87e+00	1.20e+00	1.61e+02	4.46e+00
6HKC	503	5760	1.51e-03	1.50e-03	1.00e+01	7.32e-01	1.64e+02	5.96e+00
1A91	530	8013	1.39e-03	1.29e-03	1.06e+01	2.09e+00	1.63e+02	9.78e+00
1B4R	534	7095	1.31e-03	1.12e-03	1.01e+01	1.12e+00	1.62e+02	3.91e+00
2JS9	555	8244	8.65e-04	7.86e-04	1.00e+01	1.50e+00	1.66e+02	3.93e+00
1SXL	656	8405	1.10e-03	1.26e-03	8.49e+00	1.49e+00	1.58e+02	4.78e+00

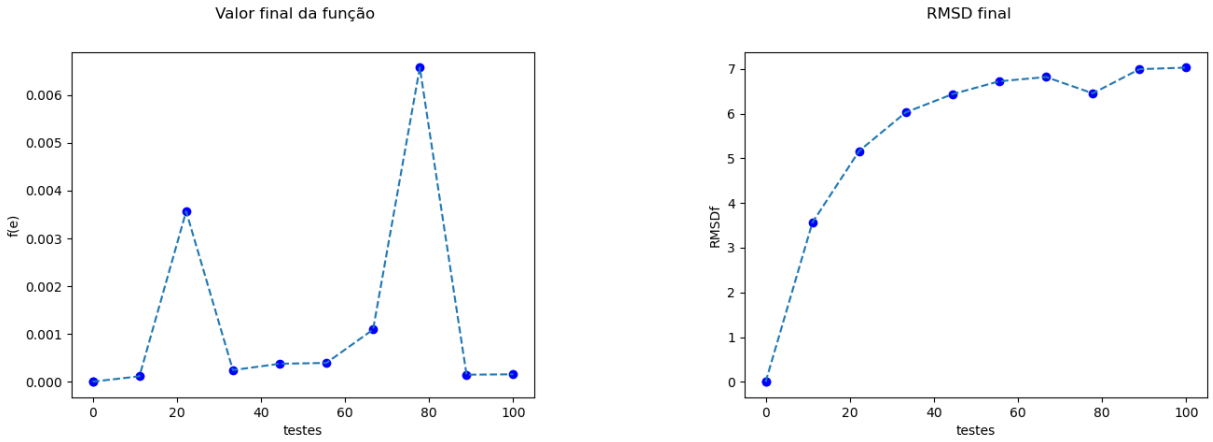
Tabela 1: Comparação de resultados entre variações de σ sob os valores de $RMSD$ avaliados nos pontos iniciais (5.4) e finais das soluções encontradas pelo SPG.

			$\sigma = 10^{-2}$		$\sigma = 10$		$\sigma = 10^{2}$	
PDB	N_a	$\|E\|$	MDE_i	MDE_f	MDE_i	MDE_f	MDE_i	MDE_f
2Y2A	30	162	8.11e-04	1.71e-05	7.09e+00	1.94e-04	7.91e+01	2.76e-04
2JMY	103	1455	2.05e-04	2.32e-05	5.67e+00	2.78e-02	6.50e+01	9.96e-03
6G4U	164	2260	2.21e-04	1.76e-05	5.65e+00	1.24e-02	6.51e+01	1.17e-02
6HN9	225	3000	1.96e-04	9.62e-06	5.87e+00	1.31e-02	6.67e+01	1.02e-02
6FS5	272	3794	1.95e-04	1.36e-05	5.45e+00	2.40e-02	6.27e+01	1.83e-02
2K35	353	4085	2.98e-04	9.62e-06	5.68e+00	4.29e-03	6.51e+01	1.27e-02
1DT4	357	3488	4.19e-04	8.03e-06	5.64e+00	9.00e-03	6.47e+01	6.24e-03
6HKC	503	5760	2.54e-04	3.64e-06	6.03e+00	2.01e-03	6.87e+01	8.12e-03
1A91	530	8013	2.03e-04	7.69e-06	5.64e+00	1.58e-02	6.47e+01	2.03e-02
1B4R	534	7095	2.27e-04	6.68e-06	5.75e+00	5.49e-03	6.57e+01	5.15e-03
2JS9	555	8244	1.88e-04	9.95e-06	5.76e+00	1.77e-02	6.61e+01	9.94e-03
1SXL	656	8405	2.25e-04	7.39e-06	5.66e+00	8.47e-03	6.50e+01	1.45e-02

Tabela 2: Comparação de resultados entre variações de σ sob os valores de MDE avaliados nos pontos iniciais (5.4) e finais das soluções encontradas pelo SPG.

a estrutura encontrada esteja um pouco distante da esperada (o que também é aceitável quando a quantidade de restrições não é suficiente para garantir a unicidade na solução). Este comportamento se torna mais evidente para a amplitude de ruído $\sigma = 10^2$.

A Figura 5.1 apresenta em mais detalhes os resultados para a proteína $2Y2A$, para variações de valor σ no intervalo $[0, 100]$.



(a) Valor funcional final

(b) RMSD

Figura 5.1: Resultados da proteína $2Y2A$ entre variações de valor σ no intervalo $[0, 100]$.

5.4 Ponto inicial da relaxação convexa

Nos experimentos da seção anterior, vimos que mesmo usando pontos iniciais como perturbações da solução esperada não garantem que o SPG encontrará a solução original. Na prática, a situação é ainda mais complicada, pois na maioria das vezes não temos ideia de que solução esperar. Isso motiva a aplicação da Relaxação Semidefinida apresentada na Seção 4 para a determinação de pontos iniciais para o método local SPG.

Para avaliar a efetividade dos pontos iniciais gerados pela Relaxação SDP, comparamos seu desempenho com o de pontos iniciais aleatórios. De fato, consideremos para cada proteína 10 pontos iniciais $X^{(i)}$ completamente aleatórios:

$$X^{(i)} = N(0, 1).$$

A Tabela 3 compreende os resultados para estes pontos iniciais. Denotamos por $f(e), MDE_f$ e $t(s)$, respectivamente, o valor final da função objetivo avaliada no ponto, o valor final do MDE e o tempo total gasto na fase SPG. Para cada um destes, apresentamos 3 colunas com o pior, melhor e a média, respectivamente, dos valores para os 10 pontos aleatórios considerados.

PDB	$f(e)$			MDE_f			$t(s)$		
2Y2A	$5.104e{-02}$	$1.282e{-04}$	$6.272e{-03}$	$2.812e{-03}$	$1.070e{-04}$	$6.293e{-04}$	2.914	0.701	1.880
2JMY	$9.920e{+01}$	$1.889e{+01}$	$5.913e{+01}$	$4.139e{-02}$	$1.025e{-02}$	$2.622e{-02}$	72.012	27.090	46.237
6G4U	$1.422e{+02}$	$6.083e{+01}$	$9.846e{+01}$	$3.693e{-02}$	$1.764e{-02}$	$2.608e{-02}$	126.184	108.328	119.401
6HN9	$1.693e{+02}$	$1.043e{+02}$	$1.274e{+02}$	$3.344e{-02}$	$2.260e{-02}$	$2.661e{-02}$	165.303	160.027	163.191
6FS5	$2.178e{+02}$	$6.773e{+01}$	$1.292e{+02}$	$3.325e{-02}$	$1.464e{-02}$	$2.232e{-02}$	207.734	201.066	204.121
2K35	$2.027e{+02}$	$9.614e{+01}$	$1.437e{+02}$	$3.476e{-02}$	$2.065e{-02}$	$2.710e{-02}$	231.882	221.765	227.090
1DT4	$1.089e{+02}$	$5.891e{+01}$	$9.392e{+01}$	$2.585e{-02}$	$1.778e{-02}$	$2.322e{-02}$	192.841	183.957	188.038
6HKC	$1.815e{+02}$	$8.796e{+01}$	$1.272e{+02}$	$2.545e{-02}$	$1.551e{-02}$	$1.942e{-02}$	323.360	305.199	315.232
1A91	$6.857e{+02}$	$4.443e{+02}$	$5.690e{+02}$	$4.580e{-02}$	$3.287e{-02}$	$3.927e{-02}$	456.011	426.721	441.591
1B4R	$4.155e{+02}$	$1.501e{+02}$	$2.763e{+02}$	$3.676e{-02}$	$1.929e{-02}$	$2.813e{-02}$	389.965	373.997	381.306
2JS9	$6.123e{+02}$	$3.889e{+02}$	$4.947e{+02}$	$4.204e{-02}$	$2.957e{-02}$	$3.505e{-02}$	462.468	439.675	450.815
1SXL	$5.100e{+02}$	$2.413e{+02}$	$3.564e{+02}$	$3.620e{-02}$	$2.114e{-02}$	$2.825e{-02}$	463.599	442.134	453.738

Tabela 3: Comparações entre pontos iniciais, resultados mostram os valores de Pior, Melhor e Média das soluções encontradas pelo SPG a partir de 10 pontos iniciais aleatórios

.

As iterações não foram apresentadas, mas na maioria dos casos, o número máximo de 2000 iterações foi atingido. Todas as instâncias aqui testadas foram criadas obedecendo os critérios estabelecidos em 5.1 utilizando como parâmetro $\Delta = 2/3$.

Estes primeiros resultados nos fornecem informações interessantes sobre o comportamento do método proposto. Entre eles podemos salientar a influência do tamanho da proteína sobre o tempo e dificuldade de resolução do problema (1.3). Por exemplo, para a proteína $1SXL$ com 1526 átomos e 656 átomos avaliados, o tempo total foi de aproximadamente $7,7$min.

PDB	N_a/N_t	$\|E\|$	d_a	It_1	$t_1(s)$	It_2	$MDEi$	$MDEf$	$f(i)$	$f(e)$	$t_2(s)$
2Y2A	30/52	162	18	16	0.204	59	$2.96e{-02}$	$3.87e{-05}$	$3.05e{+00}$	$8.05e{-06}$	0.281
2JMY	103/282	1455	45	22	10.239	98	$7.67e{-04}$	$1.85e{-05}$	$6.45e{-02}$	$4.16e{-05}$	3.658
6G4U	164/388	2260	72	24	28.509	589	$5.60e{-02}$	$5.02e{-05}$	$1.33e{+02}$	$9.93e{-04}$	34.336
6HN9	225/496	3000	99	23	51.097	2000	$6.93e{-02}$	$1.30e{-03}$	$3.35e{+02}$	$1.60e{+00}$	156.655
6FS5	272/709	3794	117	21	75.422	1006	$5.41e{-02}$	$6.94e{-04}$	$2.21e{+02}$	$1.14e{+00}$	98.938
2K35	353/616	4085	180	19	93.499	127	$1.81e{-03}$	$1.53e{-05}$	$1.94e{+00}$	$1.13e{-04}$	12.570
1DT4	357/514	3488	219	26	98.462	2000	$3.71e{-02}$	$8.46e{-04}$	$1.47e{-02}$	$1.05e{+00}$	184.260
6HKC	503/1180	5760	225	15	166.501	2000	$4.53e{-02}$	$9.35e{-04}$	$3.15e{+02}$	$1.69e{+00}$	305.360
1A91	530/1191	8013	237	16	345.835	138	$3.83e{-03}$	$1.42e{-05}$	$1.04e{+01}$	$3.57e{-04}$	27.302
1B4R	534/1114	7095	240	15	255.718	175	$2.43e{-03}$	$9.10e{-06}$	$4.35e{+00}$	$1.77e{-04}$	31.599
2JS9	555/1264	8244	243	15	341.665	1112	$3.69e{-02}$	$3.04e{-05}$	$4.31e{+02}$	$1.48e{-03}$	236.440
1SXL	656/1526	8405	291	19	525.347	1483	$1.94e{-02}$	$2.54e{-04}$	$2.22e{+02}$	$5.71e{-01}$	328.343

Tabela 4: Comparações entre pontos iniciais, resultado da fase de Relaxação Convexa.

Por outro lado, pode-se notar a melhora nos resultados obtidos pelo SPG com pontos iniciais da relaxação convexa, como mostra a Tabela 4, cuja organização se dá do seguinte modo: a primeira coluna apresenta o nome da instância, N_t representa o número total de

átomos da proteína considerada, N_a é número de átomos considerados[6], o número de distâncias $|E|$, número de distâncias exatas d_a e It_1 representa o número de iterações realizadas pela fase de relaxação convexa (referente ao Algoritmo 2) sendo t_1 o respectivo tempo total desta.

Para a parte de refino, isso é, para o método do SPG (referente ao Algoritmo 1) temos que: It_2 representa o número de iterações realizadas pelo Gradiente Espectral Projetado, $MDEi$ o valor médio das violações para o ponto inicial, $MDEf$ o valor médio das violações para o ponto obtido pelo refino, $f(i)$ o valor da função STRESS avaliada no ponto inicial, $f(e)$ o valor da função avaliada no ponto obtido pelo refino e t_2 o tempo total gasto pelo SPG para determinar a solução refinada.

Como podemos observar, em alguns casos, a maior concentração de esforço na fase de relaxação, favorece a fase de refino (SPG), em razão da qualidade do ponto inicial.

Observamos que para as proteínas $1A91, 1B4R, 2JS9$ e $6HKC$ (que possuem em torno de 500 átomos), os resultados obtidos pelo SPG com pontos inciais da relaxação semi-definida foram bem melhores (em termos de MDE final, valor funcional final e tempo gasto $t(s)$) do que usando pontos iniciais aleatórios. Por exemplo, as medidas de qualidade para a proteína $1A91$ da Tabela 4 são bem melhores do que os da Tabela 3 em um tempo total (relaxação SDP + refinamento) menor.

As Figuras 5.2, 5.3 e 5.4 trazem em detalhes os resultados da Tabela 3 nos casos das proteínas $2Y2A$, $1DT4$ e $6HKC$. Nestes gráficos, os pontos azuis representam os resultados obtidos pelo SPG para cada um dos 10 pontos aleatórios testados, a linha amarela representa a média dos valores obtidos e em vermelho a comparação com o resultado obtido usando como ponto inicial a solução da relaxação convexa.

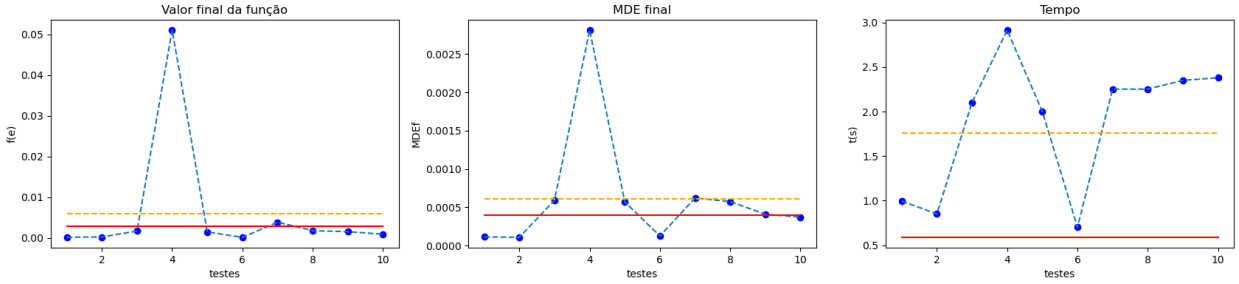


Figura 5.2: Comparações entre pontos iniciais, proteína 2Y2A

Através desses resultados torna-se fácil observar o grande potencial que a relaxação convexa nos concede em obter bons pontos iniciais para o método local SPG. As medidas de erro correspondentes ao ponto inicial proveniente da relação convexa em sua maioria sempre estão abaixo da média, quando não superam em muito os valores obtidos para pontos iniciais

[6]Aqui são considerados apenas os átomos da proteína cujas distâncias satisfazem as condições impostas sobre as instâncias artificiais

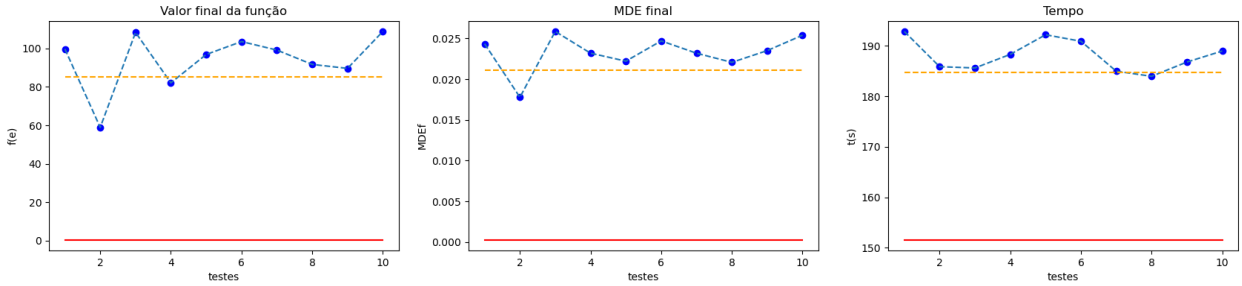


Figura 5.3: Comparações entre pontos iniciais, proteína 1DTH

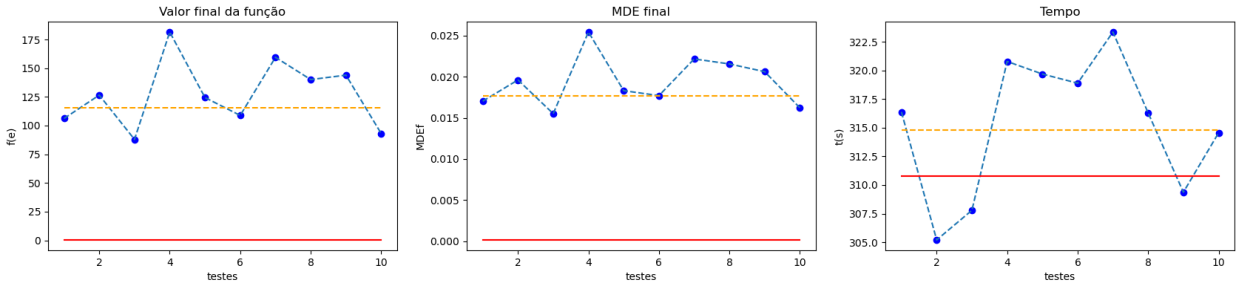


Figura 5.4: Comparações entre pontos iniciais, proteína 6HKC

aleatórios.

5.5 Comparações entre estruturas

Finalizamos este relatório com uma comparação entre as estruturas obtidas pelo esquema: relaxação SDP + refinamento com SPG, e as estruturas originais do PDB. O objetivo das Figuras 5.5, 5.6 é elucidar de maneira mais tangível os comentários que se sucederam na qualidade das soluções obtidas em ambas as fases.

Comparações entre a estrutura conhecida (em verde), com a aproximação inicial encontrada pela de relaxação convexa (em alaranjado) e com a estrutura obtida após o refino SPG (em azul). Nas Figuras 5.5 e 5.6, apresentamos uma comparação das proteínas $6HN9$ e $1SXL$ entre a estrutura original, o ponto inicial obtido da fase de programação SDP e a estrutura final do refino na fase SPG, como especificado no início desta seção. Nas Figuras 5.7 e 5.8 apresentamos também uma comparação para a proteína $2JMY$ entre a solução obtida pela relaxação convexa e a utilização de um ponto aleatório.

As imagens foram geradas utilizando dos dados obtidos pelo método discutido e o software UCSF ChimeraX 1.0 [15] para visualização e comparação das moléculas.

6 Considerações Finais

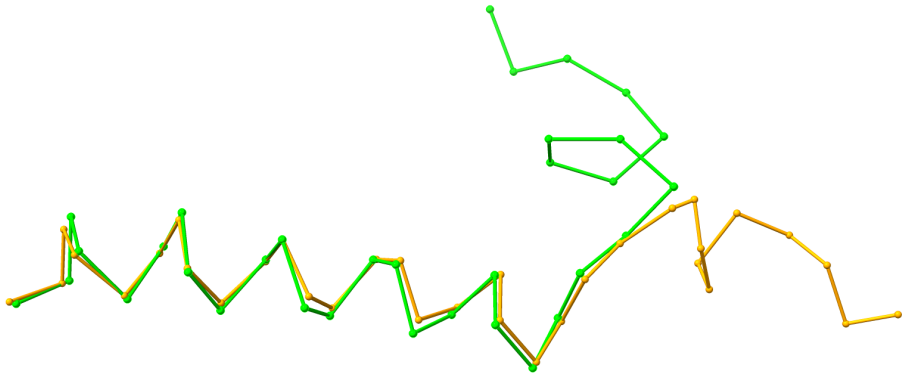
Devido ao seu carácter local, o método do Gradiente Espectral Projetado não pode garantir a obtenção de um mínimo global, e em alguns casos nem mesmo um mínimo local é obtido dentro do limite de tempo e/ou iterações, o que destaca a importância na escolha do ponto inicial.

Como pudemos observar, a estratégia de relaxação convexa para a obtenção do ponto inicial mostrou-se bem efetiva, permitindo ao método local obter uma solução de qualidade (em termos de MDE) em um tempo computacional razoável.

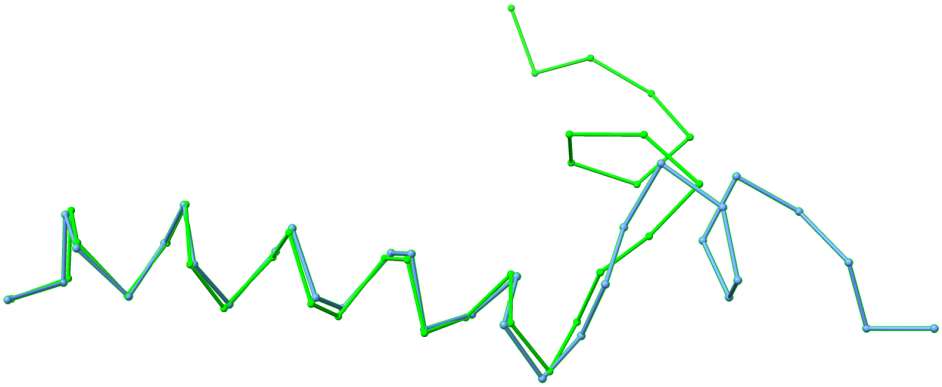
Alguns tópicos que pretendemos considerar em pesquisas futuras são: (i) experimentos com dados reais de proteínas, provenientes de laboratório; (ii) um estudo numérico comparando diferentes valores do parâmetro Δ e sua influência no desempenho do método; (iii) uma estratégia do tipo dividir-e-conquistar em que partes da proteínas são reconstruídas separadamente (em paralelo) e posteriormente alinhadas, a fim de reduzir o esforço computacional e melhorar a qualidade da estrutura final.

Agradecimentos

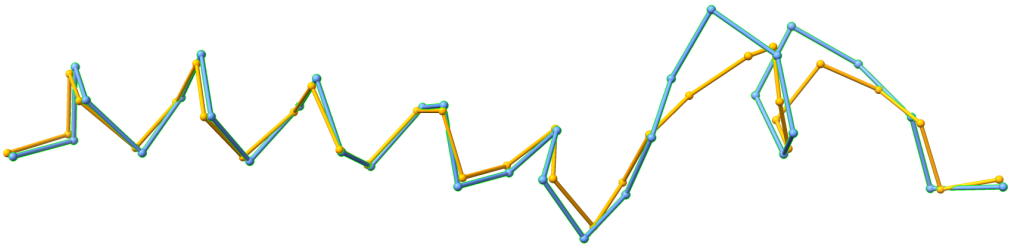
Agradecemos ao PIBIC/CNPq pela bolsa de iniciação científica no programa 2019/2020. A participação no projeto tornou-se uma experiência desafiadora e enriquecedora, permitindo que eu pudesse desenvolver minha autonomia na resolução de problemas matemáticos diversos, bem como o aprimoramento de minhas habilidades de pesquisa em referências bibliográficas e na busca por fontes variadas de informação. Gostaria de deixar claro também, que tal oportunidade salienta um inserção natural do bolsista no ambiente acadêmico, estimulando a capacidade de escrita científica, de exposição de ideias de maneira clara e concisa bem como facilitar o acesso a programas de mestrado e doutorado posteriormente.



(a) Estrutura Original (Verde) x Ponto inicial SDP (Alaranjado)

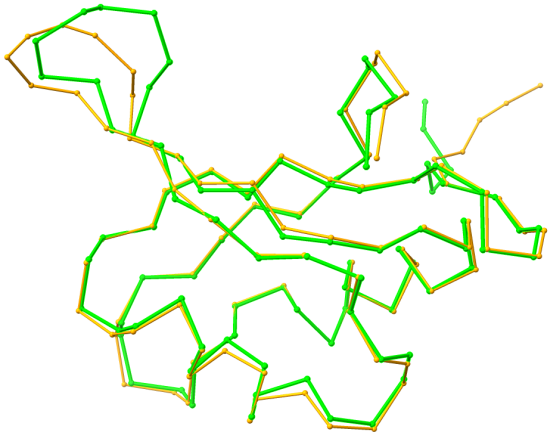


(b) Estrutura Original (Verde) x Solução refinada SPG (Azul)

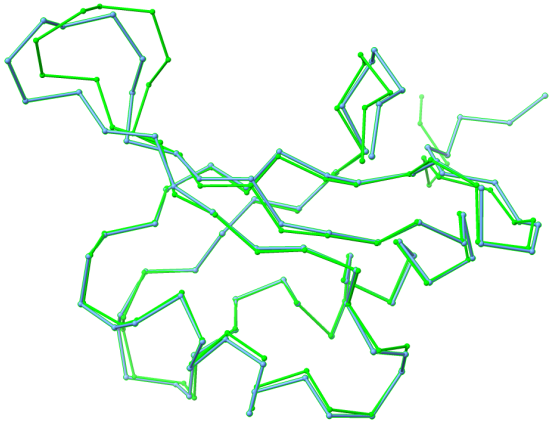


(c) Ponto inicial SDP (Alaranjado) x Solução refinada SPG (Azul)

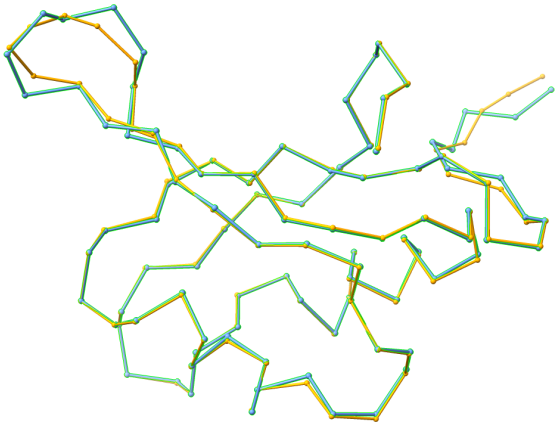
Figura 5.5: Átomos de Carbono da Cadeia Principal (6HN9)



(a) Estrutura Original (Verde) x Ponto inicial SDP (Alaranjado)



(b) Estrutura Original (Verde) x Solução refinada SPG (Azul)



(c) Ponto inicial SDP (Alaranjado) x Solução refinada SPG (Azul)

Figura 5.6: Átomos de Carbono da Cadeia Principal (1SXL)

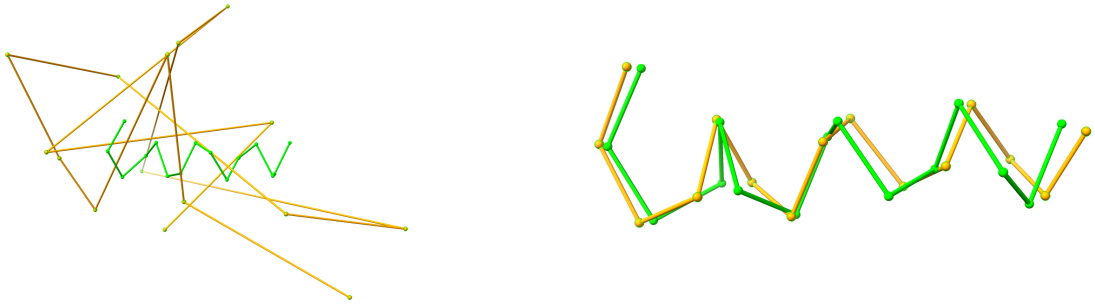


Figura 5.7: Átomos de Carbono da Cadeia Principal (2JMY). Esquerda ponto inicial aleatório (Alaranjado) à direita ponto inicial da relaxação convexa (Alaranjado).

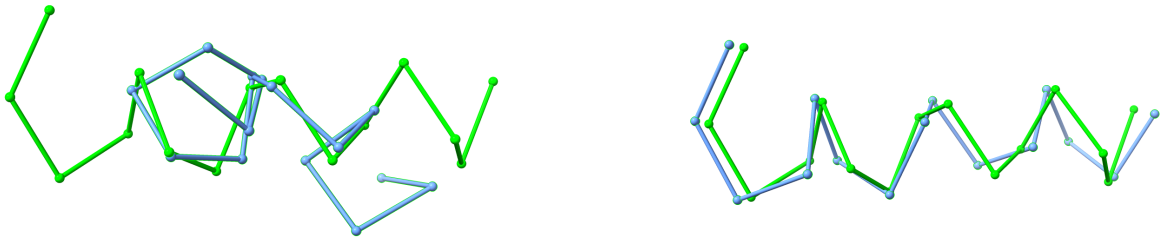


Figura 5.8: Átomos de Carbono da Cadeia Principal (2JMY). Esquerda solução do SPG com ponto inicial aleatório à direita solução do SPG com ponto inicial da relaxação convexa.

Referências

[1] S. Al-Homidan and H. Wolkowicz. Approximate and exact completion problems for euclidean distance matrices using semidefinite programming. *Linear Algebra and its Applications*, 406:109–141, 2005.

[2] B. Alipanahi, N. Krislock, A. Ghodsi, H. Wolkowicz, L. Donaldson, and M. Li. Determining protein structures from NOESY distance constraints by semidefinite programming. *Journal of Computational Biology*, 20(4):296–310, 2013.

[3] J. Barzilai and J. M. Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988.

[4] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[5] E. G. Birgin, J. M. Martínez, M. Raydan, et al. Spectral projected gradient methods: review and perspectives. *Journal of Statistical Software*, 60(3):1–21, 2014.

[6] I. Borg and P. J. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

[7] J. Cavanagh, W. J. Fairbrother, A. G. Palmer III, and N. J. Skelton. *Protein NMR spectroscopy: principles and practice*. Elsevier, 1995.

[8] V. D. Cerutti. Github repository: Protein refinement method. https://github.com/viniciusdc/Protein-Refinement, 2020.

[9] J. Dattorro. *Convex optimization & Euclidean distance geometry*. MeBoo Publishing USA, 2010.

[10] D. Di Serafino, V. Ruggiero, G. Toraldo, and L. Zanni. On the steplength selection in gradient methods for unconstrained optimization. *Applied Mathematics and Computation*, 318:176–195, 2018.

[11] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli. Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30, 2015.

[12] J. Drenth. *Principles of protein X-ray crystallography*. Springer Science & Business Media, 2007.

[13] R. Fletcher. On the Barzilai-Borwein method. In *Optimization and control with applications*, pages 235–256. Springer, 2005.

[14] W. Glunt, T. L. Hayden, and M. Raydan. Molecular conformations from distance matrices. *Journal of Computational Chemistry*, 14(1):114–120, 1993.

[15] T. D. Goddard, C. C. Huang, E. C. Meng, E. F. Pettersen, G. S. Couch, J. H. Morris, and T. E. Ferrin. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Science*, 27(1):14–25, 2018.

[16] D. S. Gonçalves. A least-squares approach for discretizable distance geometry problems with inexact distances. *Optimization Letters*, 14(2):423–437, 2020.

[17] D. S. Gonçalves, A. Mucherino, C. Lavor, and L. Liberti. Recent advances on the interval distance geometry problem. *Journal of Global optimization*, 69(3):525–545, 2017.

[18] L. Grippo, F. Lampariello, and S. Lucidi. A nonmonotone line search technique for Newton's method. *SIAM Journal on Numerical Analysis*, 23(4):707–716, 1986.

[19] P. J. Groenen, J. de Leeuw, and R. Mathar. Least squares multidimensional scaling with transformed distances. In *From data to knowledge*, pages 177–185. Springer, 1996.

[20] C. R. Johnson and P. Tarazaga. Connections between the real positive semidefinite and distance matrix completion problems. *Linear Algebra and its Applications*, 223:375–391, 1995.

[21] M. Raydan. The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. *SIAM Journal on Optimization*, 7(1):26–33, 1997.

[22] M. Raydan M. *Convergence properties of the Barzilai and Borwein gradient method*. PhD thesis, 1991.

[23] I. Schoenberg. Remarks to M. Frechet's article "Sur la definition axiomatique d'une classe d'espaces vectoriels distanci es applicables vectoriellement sur l'espace de Hilbert". *Annals of Mathematics*, 36:724–732, 1935.

[24] M. Souza, C. Lavor, A. Muritiba, and N. Maculan. Solving the molecular distance geometry problem with inaccurate distance data. *BMC Bioinformatics*, 14(S9):S7, 2013.

[25] W. Sun and Y.-X. Yuan. *Optimization theory and methods: nonlinear programming*, volume 1. Springer Science & Business Media, 2006.

[26] R. H. Tütüncü, K.-C. Toh, and M. J. Todd. Solving semidefinite-quadratic-linear programs using SDPT3. *Mathematical programming*, 95(2):189–217, 2003.

[27] K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.

[28] K. Wüthrich. NMR with proteins and nucleic acids. *Europhysics News*, 17(1):11–13, 1986.

[29] G. Young and A. S. Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3(1):19–22, 1938.