Trabalho apresentado no XXXVII CNMAC, S.J. dos Campos - SP, 2017.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

O problema da distância geométrica intervalar via otimização global

Luiz Leduino de Salles Neto[1]
Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo, São Jose dos Campos, SP
Weldon Lodwick[2]
Department of Mathematical and Statistical Sciences, University of Colorado Denver, Denver, CO, USA

Resumo. Neste trabalho apresentamos uma nova modelagem matemática, via otimização global e funcção intervalar, para o problema da distância geométrica intervalar (PDGi). O objetivo do PDGi, no contexto da estrutura de proteínas, consiste em encontrar uma realização em R^3 de um grafo G=(V,E), onde as distancias entre os vértices sao dadas por intervalos, em conformidade com as medidas experimentais obtidas pela Ressonância Magnética Nuclear. Em particular, esse trabalho aborda o problema de encontrar a posição de cada átomo (vértice) de uma proteína, dado as distâncias intervalares desse vértice a três vértices anteriores. Os resultados demonstram que a abordagem é promissora.

Palavras-chave. Problema da Distância Geométrica, Otimização , Incerteza

1 Introdução

O problema da distância geométrica (PDG) tem sido estudado por diversos pesquisadores, possuindo aplicações na determinação da geometria de proteinas, na localização de sensores, com aplicações na internet das coisas, robotica, entre outros [4, 5].

Em especial é possivel, por meio de técnicas como a Ressonância Magnética Nuclear (RMN), estimar distâncias entre pares de átomos de uma determinada molécula, e o problema se torna o de identificar a conformação tridimensional da molécula, isto é, as posições de todos os seus átomos. Neste campo, o principal interesse é sobre as proteínas, visto que a descoberta de sua conformação tridimensional nos permite obter informações sobre as funções que as proteínas são capazes de realizar. Logo, é possível concluir que o PDG tem implicações no desenvolvimento de novos fármacos e uma vasta aplicação em biotecnologia. Quando se trata de moléculas biológicas, o PDG é geralmente conhecido como PDG molecular (PDGM).

O PDG em R^3 pode ser definido, de forma geral, da seguinte forma:

Definição 1.1. *Dado um grafo $G = (V, E, d)$ com pesos $d : E \to (0, \infty)$ em suas arestas o PDG consiste em encontrar uma função $x : V \to R^3$ tal que $||x(u) - x(v)||_2 = d(u, v)$ para*

[1]luiz.leduino@unifesp.br
[2]Weldon.Lodwick@ucdenver.edu

DOI: 10.5540/03.2018.006.01.0408

todo $u, v \in V$. A solução é, dessa forma, associar a cada vértice de G um único ponto em R^3 satisfazendo as equações acima.

Observação 1.1. *Nesse trabalho utilizamos apenas a distância euclidiana. Assim, se $u = (u_1, u_2, u_3)$ e $v = (v_1, v_2, v_3) \in R^3$, temos que:*

$$||u - v||_2 = d(u, v) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + (u_3 - v_3)^2}$$

Uma forma de abordar o PDG é resolver o seguinte problema de otimização :

Definição 1.2. *Dado um grafo G=(V,E,D) associamos o problema de otimização da distância geométrica:*

$$\begin{array}{rl} \textit{(PODG)} \quad \textit{Min} & \sum ||x(u) - x(v)||_2^2 - d(u, v)^2)^2 \\ & x(u) \in R^3, u, v \in V, \end{array}$$

Em muitas situações , como no problema da geometria de proteínas, em virtude de limitações experimentais da Ressonancia Magnetica Nuclear (RMN) algumas distâncias são dadas como intervalos. Nesse caso definirmos o Problema da Distância Geométrica Intervalar (PDGi) da seguinte forma:

Definição 1.3. *Dado um grafo $G = (V, E, [di, ds])$ com uma distância-intervalar $[di, ds]$ associada a cada uma de suas arestas, sendo $di(v_l, v_k) \leq ds(v_l, v_k)$, o PDGi consiste em encontrar uma funcao $x : V \to R^3$ tal que $di(v_l, v_k) \leq ||x(v_l) - x(v_k)|| \leq ds(v_l, v_k)$ para todo $v_l, v_k \in V$. Nós assumimos que:*

1. *São dadas as coordenadas dos três primeiros vértices v_1, v_2, v_3;*
2. *Para todo $i = 4, 5, ..., n$ os três vèrtices anteriores formam um clique com v_i, isto è, $\{v_{i-3}, v_{i-2}, v_{i-1}, v_i\}$ è um clique onde:*

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

3. *Para algum $i = 5, \ldots n$ existe $j \in \{1, 2, \ldots, i - 4\}$ tal que:*

$$[d(l, k)] = [di_{v_l,v_k}, ds_{v_l,v_k}],\ 0 < di_{v_l,v_k} < ds_{v_l,v_k} \tag{1}$$

Em [1] os autores apresentam um conjunto de formulações de programação quadrática e de programação semidefinida bem como um conjunto de métodos novos e existentes para resolver estas formulações . Entretanto, a abordagem proposta nesse trabalho difere de todas essas formulações , e consequentemente, de seus métodos de resolução . Os testes realizados indicam que a abordagem proposta nesse trabalho é promissora quando comparado com os melhores resultados obtidos em [1].

Entre todas as formulações propostas em [1] a que obteve melhores resultados foi a seguinte:

DOI: 10.5540/03.2018.006.01.0408

$$\begin{gathered}\text{Min } \sum_{v_k,v_l \in E} s_{v_k v_l} \\ \text{s. a: } di^2_{v_k,v_l} - ||x_{v_k} - x_{v_l}||_2^2 \leq s_{v_k,v_l} \\ ||x_{v_k} - x_{v_l}||_2^2 - di^2_{v_k,v_l} \leq s_{v_k,v_l} \\ \sum_{v \in V} x_{vm} = 0 \ \forall \ m \leq 3\end{gathered}$$

Já o método que obteve melhores resultados em [1] foi a VNS - Variable Neighborhood Search. A VNS é uma metaheuristica baseada em mudanÃ§as sistematicas na estrutura de vizinhanca dentro de uma busca, visando escapar de otimos locais [6].

Na próxima seção apresentamos o citado novo modelo matemático para o PDGi via otimização global. Na seção 3 apresentamos alguns resultados dos testes computacionais. Por fim, na seção 4 apresentamos nossas conclusões e perspectivas.

2 Modelo matematico para o PDGi

Para a obtenção de uma solução para o PDGi cada distância intervalar entre dois vértices foi considerada como uma função linear. Mais especificamente, para todo $(v_k, v_l) \in E$ definimos uma função associada da seguinte forma:

Definição 2.1. *A cada intervalo* $[di(v_k, v_l), ds(v_k, v_l)]$ *associamos a função :*

$$f_{kl}(\lambda) = di(v_k, v_l) + \lambda * (ds(v_k, v_l) - di(v_k, v_l)) \tag{2}$$

onde $\lambda \in [0, 1]$.

As propriedades dessa função , bem como os principais conceitos da aritmética intervalar podem ser encontrados em [7].

Utilizando a função f_{kl} (2) para representar a distância intervalar, e considerando as hipóteses da definição 1.3, será preciso resolver $|V| - 3$ sub-problemas de minimização da seguinte forma:

$$\begin{gathered}\text{Min } \sum_{l=1}^{3} [(x(1) - y(k-l,1))^2 + (x(2) - y(k-l,2))^2 + (x(3) - y(l-3,3)^2 - f_{kl}(\lambda_l)^2]^2 \\ \text{s. a: } 0 \leq \lambda_l \leq 1 \text{ para } l = 1, 2, 3. \\ x \in R^3\end{gathered}$$

onde $l = 4, \ldots, |V|$.

Assim, temos $|V| - 3$ sub-problemas de otimização , com até 6 variáveis, onde em cada problema encontramos as coordenadas de um vértice/atomo, que serão utilizadas no sub-problema seguinte para a obtenção das coordenadas do vértice corrente, ate a obtenção do posicionamento de todos os vértices e, assim, a solução do PDGi.

3 Implementação e resultados computacionais

Para os testes computacionais da abordagem proposta foi implementado um algoritmo utilizando a função *optim* do software livre *scilab* [2]. Dentre as opções de método dessa função nesse trabalho foi utilizada o método quase newton [3].

O algoritmo foi rodado em um laptop HP, 4GB RAM, processador intel celeron 1.6 GHz, 64 bit, sistema operacional Windows Home 10.

Para buscar o ótimo global de cada um dos sub-problemas de otimização foram geradas 50 soluções iniciais aleatórias se o sub-problema possuia um intervalo não degenerado, ou seja, onde $d_i < d_s$ para algum par de vértices presentes no sub-problema corrente. Caso o sub-problema não possuisse intervalos não degenerados, ou seja, todos os intervalos seriam números reais ($d_i = d_s$), foram geradas 20 soluções iniciais aleatórias.

Para aferir a qualidade da solução foi utilizada a métrica $\psi(x, G)$ proposta em [1]. Sejam:

- $$\alpha_{v_l v_k}(x) = max(0, di(v_l, v_k) - ||x_{v_l} - x_{v_k}||_2) + max(0, ||x_{v_l} - x_{v_k}||_2 - ds(v_l, v_k)) \quad (3)$$

  se $di(v_l, v_k) \neq ds(v_l, v_k)$.

- $$\alpha_{v_l v_k}(x) = |||x_{v_l} - x_{v_k}||_2 - d(v_l, v_k) \quad (4)$$

  se $di(v_l, v_k) = ds(v_l, v_k) = d(v_l, v_k)$.

- $$\psi(x, G) = \max\{\alpha_{v_l v_k}(x)\} \quad (5)$$

Assim, essa métrica retorna a pior diferença entre a solução encontrada e as distâncias intervalares dadas.

Para realizar os testes computacionais foram utilizados os dados de proteínas presentes no banco de dados PDB - *Protein Data Bank* [8], o mais importante banco de dados de proteínas utilizado por pesquisadores de diversas áreas em todo mundo. Nesse banco de dados é possivel obter as coordenadas de diversas proteínas. Para obter as distâncias entre os átomos foi realizado um pré-processamento.

Visando simular os dados obtidos por RMN foram utilizadas apenas as distâncias entre átomos, iguais, de hidrogênio ou carbono.

Para a escolha de quais distâncias $d(v_l, v_k)$ seriam, de fato, intervalares foi gerado um número pseudo-aleatório δ para cada aresta do grafo, utilizando-se a função *ran()* do *scilab*. Verifica-se, então, se $\delta < 0.1$. Se sim, foi define-se $di(v_l, v_k) = d(v_l, v_k) - 0.1$ e $ds(v_l, v_k) = d(v_l, v_k) + 0.1$. Caso contrário a distância associada a aresta nao é modificada.

As proteínas, suas características e os resultados obtidos estão descritos abaixo:

- **Proteína 1BRV** - proteína imunodominante do vírus respiratório sincicial bovino, possui 130 átomos de hidrogênio. Entre todas as distâncias calculadas 417 se constituíram como intervalares pela metodologia acima exposta. Para os cálculos do posicionamento dos 127 átomos, do total de 130, já que o posicionamento dos 3 primeiros átomos e um parâmetro de entrada, por meio de 127 sub-problemas de otimização , em 22 deles havia pelo menos uma distância intervalar.

  Melhor solução geral encontrada: $\psi(x, G) = 1.13$. Tempo computacional = 14.72 segundos.

DOI: 10.5540/03.2018.006.01.0408

- **Proteína 100D** - estrutura de cristal, possui 200 átomos de carbono. Entre todas as distâncias calculadas 960 se constitiuíram como intervalares pela metdologia exposta acima. Para os cálculos do posicionamento dos 197 átomos, do total de 200, por meio de 197 sub-problemas de otimização , em 24 deles havia pelo menos uma distância intervalar.

  Melhor solução geral encontrada: $\psi(x, G) = 0.615$. Tempo computacional = 25.02 segundos.

- **Proteina 1PPT** - polipetidio pancreático, possui 192 átomos de carbono. Entre todas as distâncias calculadas 869 se constitiuíram como intervalares pela metdologia exposta acima. Para os cálculos do posicionamento dos 189 átomos, do total de 192, por meio de 189 sub-problemas de otimização , em 28 deles havia pelo menos uma distância intervalar.

  Melhor solução geral encontrada: $\psi(x, G) = 2.99$. Tempo computacional = 20.93 segundos.

Vale observar que em [1] o melhor resultado obtido entre as diversas combinações de formulações matemáticas e métodos computacionais para a Proteína 100D foi $\psi(x, G) = 2.05$, e para a proteína 1PPT foi $\psi(x, G) = 1.93$. Contudo, é importante ressaltar que não é possível dizer qual abordagem é melhor, se a abordagem proposta nesse trabalho ou as melhores combinações (formulação + método) propostas em [1], pois as distâncias intervalares não são as mesmas, logo, não ''e possível comparar os métodos. De qualquer forma, é possivel concluir que a abordagem proposta nesse trabalho é, pelo menos, promissora.

4 Conclusões

Nesse trabalho é apresentado uma nova abordagem do problema da distância geométrica intervalar (PDGi). Para modelar o PDGi como um problema de otimização global irrestrita foi utilizado o conceito de função intervalar. Os testes realizados com dados reais de proteínas, introduzindo a distância intervalar não degenrada em cerca de 10% das distâncias avaliadas, demonstram que a abordagem é promissora e deve ter seu estudo e testes aprofundados.

Referências

[1] C. D'Ambrosio, V. Ky, C. Lavor, L. Liberti, N. Maculan, *Solving distance geometry problems with interval data using formulation-based methods.* Technical report, LIX Ecole Polytechnique, 2014.

[2] M. Baudin, S. Steer, Optimization with scilab, present and future. In *Open-source Software for Scientific Computation* (OSSC), IEEE International Workshop, pages 99-106, 2009.

[3] M. Baudin, V. Couvert, S. Steer, Optimization in scilab. Scilab Consortium, INRIA Paris-Rocquencourt, July 2010.

[4] C. Lavor, L. Liberti, A. Mucherino, The interval Branch-and-Prune algorithm for the discretizable molecular distance geometry problem with inexact distances, *Journal of Global Optimization* 56: 855-871, 2013.

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

[6] M. Mladenovic, P. Hansen, Variable neighborhood search. *Computers & operations research*, 24: 1097-1100, 1997.

[7] R. Moore, W. Lodwick, Interval analysis and fuzzy set theory. *Fuzzy sets and systems*, 135: 5-9, 2003.

[8] www.rcsb.org/pdb Acessado em 27 de marco de 2017.

