Um estudo teórico-computacional de uma aplicação da Geometria de Distâncias em conformações proteicas

Guilherme Philippi[1] Felipe Fidalgo[2]

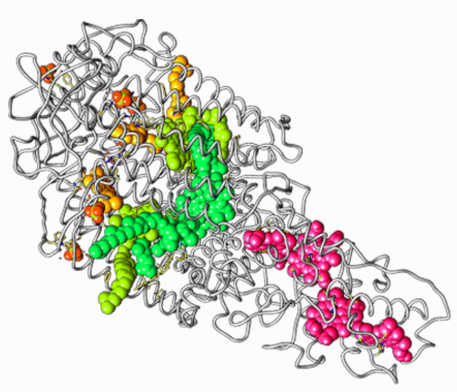
g.philippi@grad.ufsc.br[1], felipe.fidalgo@ufsc.br[2]

UNIVERSIDADE FEDERAL DE SANTA CATARINA

CNMAC 2021
XL Congresso Nacional de Matemática Aplicada e Computacional

Motivação

Existe uma relação muito forte com a forma geométrica das moléculas orgânicas e suas funções em organismos vivos [1]. Sabe-se que uma das estruturas principais da vida é construída com os aminoácidos que formam as proteínas. Logo, para conhecer a estrutura geométrica dessas moléculas, é preciso estudar a sua geometria [2]. O *Worldwide Protein Data Bank* (ou wwPDB) é um repositório online contendo os dados de todas as proteínas já catalogadas.



Geometria de Distâncias

A Geometria de Distâncias originou-se dos esforços de Menger (1928) [3], seguido por Blumenthal (1953) [4], ao caracterizar vários conceitos da Geometria Euclidiana (como congruência e convexidade) em termos de distâncias [2]. O desafio fundamental dessa área é o estudo de um problema inverso denominado Problema de Geometria de Distâncias (do inglês, DGP), onde, dados um grafo simples, ponderado positivamente e não direcionado $G = (V, E, d)$ e um inteiro $K > 0$, deseja-se encontrar uma imersão $x : V \to \mathbb{R}^K$ (a qual é chamada de realização de G em $\mathbb{R}^K$) tal que

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

Em particular, a restrição do DGP para $k = 3$ é de interesse prático e conhecido como Problema de Geometria de Distâncias Moleculares (do inglês, MDGP), pois surgiu na busca por conformações moleculares tridimensionais [2].

DMDGP

Para que se crie um ambiente adequado para encontrar conformações, especificamente, para proteínas, uma relação de ordem total no conjunto V pode ser encontrada. Munido de tal ordem, o espaço de busca por soluções do MDGP pode ser discretizado [5].

Discretizable MDGP: Dados um grafo ponderado e não-direcionado $G = (V, E, d)$, onde $d : E \longrightarrow \mathbb{R}_+$, o subconjunto de vértices iniciais $U_0 = \{v_1, v_2, v_3\}$ e uma relação de ordem total em V que satisfaz a seguinte relação de axiomas:

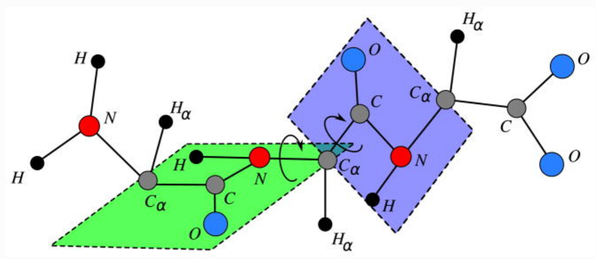
1. U_0 é um 3-clique em G (inicialização);
2. $\forall\, v_i$ tal que $i > 3$ nessa ordem, $U_i = \{v_i, v_{i-1}, v_{i-2}, v_{i-3}\}$ é um 4-clique em G (hipótese de discretização);
3. $\forall\, v_i$ tal que $i > 3$ nessa ordem, juntamente com $\{v_{i-3}, v_{i-2}, v_{i-1}\}$, vale a desigualdade

$$d_{i-3,i-1} < d_{i-3,i-2} + d_{i-2,i-1},$$
(Desigualdade Triangular Estrita)

encontre uma imersão $x : V \longrightarrow \mathbb{R}^3$ tal que valha $\|x(v_i) - x(v_j)\| = d_{i,j}, \forall \{v_i, v_j\} \in E$.

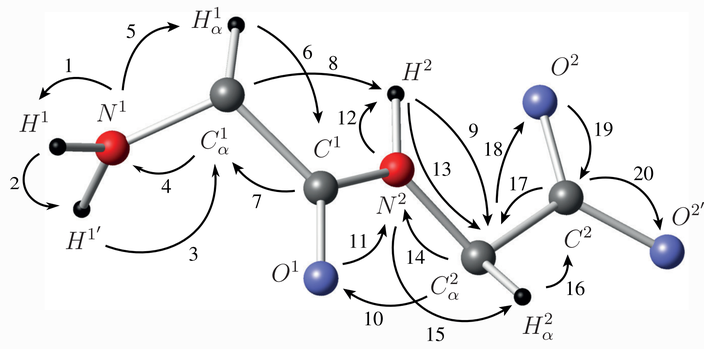
Geometria das Proteínas

Para ser possível encontrar a ordem acima, precisamos estudar a geometria molecular. Felizmente existe uma subestrutura periódica nas proteínas chamada **Cadeia Principal** (ou, *backbone*), que possui uma geometria rica e bem conhecida [1]. Através de dados de cristalografia, pode-se estimar distâncias entre pares de átomos ligados por ligações covalentes e ângulos entre três átomos separados por duas ligações covalentes. Além disso, há um plano formado por ligação peptídicas nesta estrutura [6].



Ordem Conveniente

Tendo posse dessas informações, pode-se pensar em percorrer os átomos da molécula utilizando esta subestrutura como guia, repetindo átomos, afim de fazer valer os três axiomas do DMDGP. Isto foi feito em [6] propondo o ***hand-crafted vertex order***, conforme esboçada abaixo.



HC Order

Seja $G = (V, E, d)$ o grafo associado a cadeia principal de uma proteína ($\{N^k, C^k_\alpha, C^k\}$, para $k = 1, \ldots, p$), incluindo os átomos de oxigênio O^k, ligados ao C^k, e átomos de hidrogênio H^k e H^k_α, ligados ao N^k e C^k_α, respectivamente (conforme imagem acima, onde $p = 3$).
Define-se a ordem HC como:

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

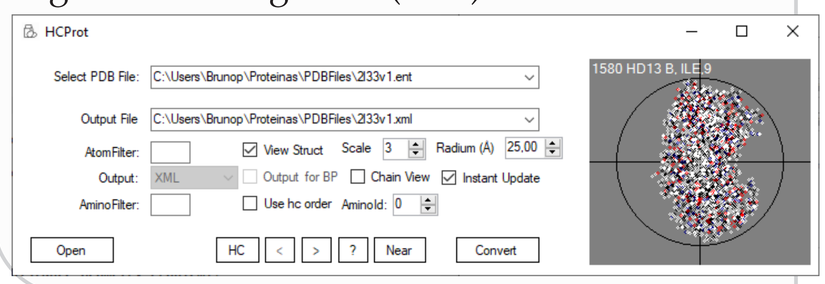
Onde, como na figura, $i = 2, \ldots, p-1$, $H^{1'}$ é o segundo hidrogênio ligado ao N^1 e $O^{p'}$ é o segundo oxigênio ligado ao C^p.

Softare HCProt

Para facilitar o estudo da geometria molecular e as simulações do problema, implementou-se um software chamado **HCProt**, que aceita como entrada arquivos do repositório wwPDB e tem como saída um arquivo descrevendo a proteína reordenada (utilizando, por exemplo, a ordenação HC). O software possui duas versões: uma contendo interface gráfica, que permite a visualização de uma projeção 2D da proteína mas é limitada ao SO Windows, e outra com interface CLI (chamada **HCProtCLI**) que é multiplataforma. Ambas podem ser encontradas em repositórios públicos no GitHub.

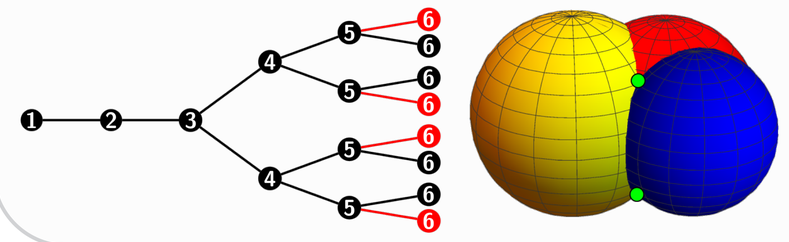
Inferface HCProt

Segue interface gráfica (GUI) do software.



Algorítimo Branch-&-Prune

Ganhamos algumas vantagens com a discretização do problema pois a ordem no DMDGP garante a finitude do conjunto solução do problema e, além disso, organiza o espaço onde devemos fazer a busca por uma solução. Na verdade, a ordem induz uma estrutura de **árvore binária** no espaço de busca [5]. De fato, a partir do quarto, sempre temos no máximo duas possibilidades para posicionar o próximo vértice. Devido a esta estrutura, criou-se o algorítimo ***Branch-&-Prune***, que consiste em uma estratégia numérica recursiva que resolve o DMDGP eficientemente utilizando uma busca combinatória no espaço de busca por soluções, onde realiza-se vértice por vértice do sistema, seguindo a ordem dada, "podando" todo sub-conjunto solução infactível do sistema em relação a distâcnias extras do grafo.



Referências

[1] Nelson, D. L. and Cox, M. M. *Lehninger principles of biochemistry, 6th edition*. W.H.Freeman and Company, New York, 2012.

[2] Liberti, L., Lavor, C., Maculan, N., e Mucherino, A. (2014). Euclidean distance geometry and applications. *SIAM review*, 56:3-69. DOI:10.1137/120875909

[3] Menger, K. Untersuchungen über allgemeine Metrik, *Math. Ann.*, 100:75–163, 1928. DOI:doi.org/10.1007/BF01448840.

[4] Blumenthal, L. M. *Theory and applications of distance geometry*. Oxford University Press, Oxford, 1953. conveniente

[5] Lavor, C., Liberti, L., Maculan, N., and Mucherino, A. The discretizable molecular distance geometry problem, *Computational Optimization and Application*, Springer, volume 52, number 1, pages 115-146, 2012. DOI. 10.1007/s10589-011-9402-6.

[6] Lavor, C., Liberti, L., Donald, B., Worley, B., Bardiaux, B., Malliavin, T. E. and Nilges, M. Minimal NMR distance information for rigidity of protein graphs, *Discrete Applied Mathematics*, Elsevier, 256:91–104, 2019. DOI:10.1016/j.dam.2018.03.071.

Agradecimentos

O presente trabalho foi realizado com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq – Brasil. Agradecemos a organização do evento.