# DMDGP: Um Problema Real

Guilherme Philippi  Felipe Fidalgo

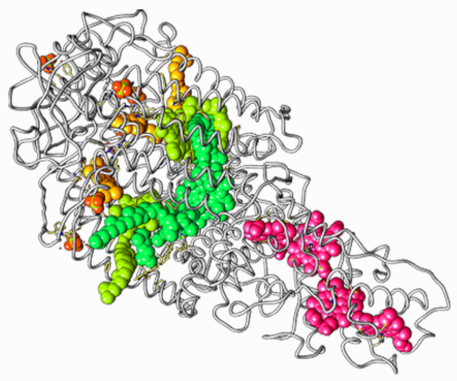
g.philippi@grad.ufsc.br, felipe.fidalgo@ufsc.br

UNIVERSIDADE FEDERAL DE SANTA CATARINA

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico

## 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]. Existe uma ferramenta muito importante para quem trabalha nesta área: um repositório online contendo dados de todas as proteínas catalogadas chamado *Worldwide Protein Data Bank* (ou wwPDB).



## O Problema

Abordaremos aqui o problema de **encontrar as posições dos átomos de uma molécula**, tendo como entradas algumas distâncias entre átomos próximos (obtidas através de experimentos de Ressonância Magnética Nuclear [3]). Este é conhecido na literatura como *Molecular Distance Geometry Problem* (MDGP), que é uma particularização do *Distance Geometry Problem* (DGP) [4]. Tal problema, **munido de uma ordem conveniente** para percorrer seus átomos (dada pelo *Discretization Vertex Order Problem*, ou simplesmente, DVOP), **pode ser discretizado**, gerando o *Discretizable MDGP* (DMDGP), como segue formalmente definido [5].

## DMDGP

**Discretizable Molecular Distance Geometry Problem:** 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. $G[U_0]$ é um clique em três vértices (iniciando a configuração);
2. para todo vértice $v_i$ com posto $i = \rho(v_i) > 3$ nesta ordem, $G[U_i]$ é uma clique com quatro vértices (ordem de discretização, dada anteriormente) e
3. para cada vértice $v_i$, com posto $i = \rho(v_i) > 3$, 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)

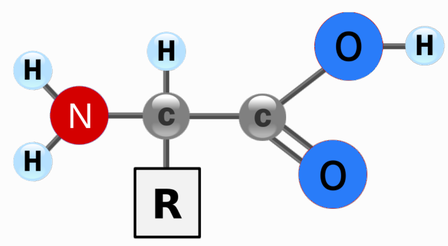
encontre uma imersão $x : V \longrightarrow \mathbb{R}^3$ tal que valha $\|x(v_i) - x(v_j)\| = d_{i,j}, \forall \{v_i, v_j\} \in E$.

## 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** [1].

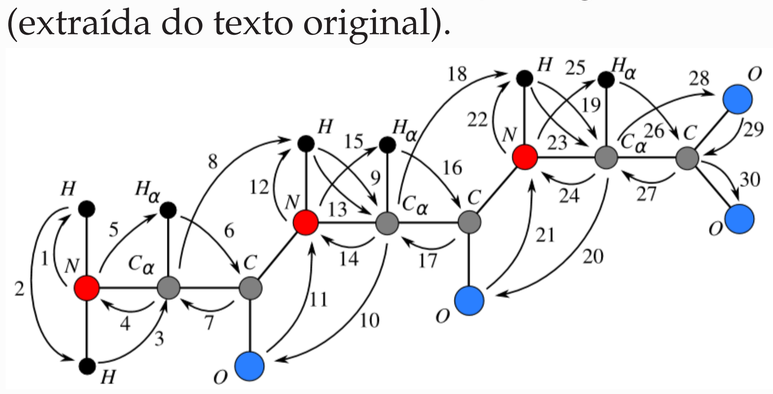
## Cadeia Principal

Através de dados experimentais de cristalografia, **sabe-se muito sobre a geometria média dessa subestrutura** [6] (também conhecida como *backbone*, mostrada abaixo) onde **os comprimentos e ângulos** entre as ligações dos átomos que a formam **são fixos**, na média, a menos de erros de medida.




## Ordem Conveniente

Tendo posse dessas informações sobre a cadeia principal, pode-se pensar em percorrer os átomos da molécula **utilizando esta subestrutura como guia**, *repetindo* os átomos que possuem propriedades conhecidas, afim de fazer valer os três axiomas do DMDGP. Isto foi feito em [5] propondo o ***hand-crafted vertex order***, conforme esboça a figura abaixo (extraída do texto original).



## 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 $H^k_\alpha$, ligados ao $N^k$ e $C^k_\alpha$, respectivamente (conforme imagem acima, onde $p = 3$).
Define-se a ordem HC como:

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

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 $C^p$.

## PDBReader

Para facilitar o estudo da geometria molecular e as simulações do problema, o autor deste documento implementou um software (chamado **PDBReader**) que aceita como entrada instâncias de proteínas do repositório wwPDB e tem como saída um arquivo descrevendo a proteína **com sua instância reordenada** (utilizando ordenação HC). Perceba que esta é uma *automação da discretização do problema*.

## Results

- $1^{st}$ **column**: name of the PDB protein;
- $2^{nd}$ **column**: number of atoms;
- $3^{rd}$ **column**: RMSD values for UT;
- $4^{th}$ **column**: RMSD values for UGB;
- $5^{th}$ **column**: relative time value UT/UGB;

| Name | # atoms | UT | UGB | Time(UT/UGB) |
|---|---|---|---|---|
| 1ID7 | 189 | 3,12E-09 | 3,12E-09 | 57,32% |
| 1FW5 | 332 | 1,18E-08 | 1,18E-08 | 60,42% |
| 1JAV | 360 | 1,38E-07 | 1,38E-07 | 60,70% |
| 1MEQ | 405 | 6,39E-11 | 6,42E-11 | 61,51% |
| 1AMB | 438 | 8,22E-06 | 8,22E-06 | 61,15% |
| 1R7C | 532 | 8,39E-07 | 8,38E-07 | 63,83% |
| 1HLL | 540 | 1,59E-07 | 1,59E-07 | 61,96% |
| 1VII | 596 | 9,19E-07 | 9,19E-07 | 58,55% |
| 1HIP | 617 | 3,15E-09 | 3,14E-09 | 57,98% |
| 1ULR | 677 | 2,82E-09 | 2,81E-09 | 59,20% |
| 1KVX | 954 | 1,65E-06 | 1,65E-06 | 58,19% |
| 1VMP | 1166 | 1,97E-07 | 1,97E-07 | 56,25% |
| 1RGS | 2015 | 1,37E-08 | 1,37E-08 | 55,23% |
| 1BPM | 3671 | 5,08E-06 | 5,08E-06 | 57,11% |

## Concluding Remarks

UT has shown good numerical stability and accuracy, when compared to one method from the literature, being better conditioned and requiring less time to solve the MDGP. Even the difference is not much, it can play an important role when $n$ grows.

## Future Work

- Our main outlook: treat noisy distances, i. e., sparse and inexact ones;
- Stochastic modeling is one of the ideas, by means of the Monte Carlo approach;
- Control even more the growing of the condition number of the linear system;
- Identify $t_j$ with such modeling of uncertainties with respect to the MDGP in the sense of Monte Carlo simulations

## References

[1] David L Nelson and Michael M Cox. *Lehninger: principios de bioquímica*. 2015.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

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

## Acknoledgments

We would like to thank to the brazilian agencies CNPq