SMNE UFPR
Simpósio de Métodos Numéricos em Engenharia
25 a 27 de outubro, 2017

Reconstrução de Estruturas Proteicas via Otimização Contínua

Alisson Lucas de Souza*† and André Luís Machado Martinez *‡
*Departamento Acadêmico de Matemática
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Brasil
† Email: alissonsouza@alunos.utfpr.edu.br
‡ Email: martinez@utfpr.edu.br

***Resumo*—Neste trabalho apresenta-se o Problema da Geometria de Distâncias Moleculares (PGDM) para reconstrução de estruturas tridimensionais de proteínas por meio de métodos de Otimização Contínua. O objetivo do trabalho é analisar o desempenho dos métodos clássicos de Otimização Não-Linear (método BFGS) e de Mínimos-Quadrados (método Gauss-Newton) na resolução do PGDM, cujos dados estruturais das proteínas utilizadas no trabalho foram resgatados do Protein Data Bank (PDB). Apresenta-se e discute-se os resultados obtidos, além de propor-se novas modificações como perspectivas futuras.**

***Palavras-chave*—Problema da Geometria de Distâncias Moleculares; Otimização Contínua; Proteínas; BFGS; Gauss-Newton.**

I. Introdução

Neste trabalho apresentamos um estudo contínuo e discreto sobre reconstrução de estruturas tridimensionais de proteínas. Proteínas são compostos orgânicos constituídos de aminoácidos que estão conectados entre si por ligações peptídicas. Em geral, os aminoácidos são moléculas contendo os grupos funcionais amina ($COOH$) e carboxila ($NH2$), presos ao mesmo átomo de carbono, também conhecido como carbono alfa (C_α).

O estudo da estrutura tridimensional de uma proteína é fundamental para o conhecimento do papel que ela desempenha nos organismos. Além disso, como a quantidade de proteínas conhecidas é muito grande, faz-se necessária a utilização de um banco de dados para armazenamento das informações estruturais e o principal banco de dados é conhecido como Protein Data Bank (PDB).

Para realizar a tarefa de reconstrução das proteínas, este trabalho baseia-se na resolução do Problema da Geometria de Distâncias Moleculares (PGDM) via Otimização Contínua, comparando os resultados computacionais obtidos utilizando os métodos de Gauss-Newton e BFGS. Nesta abordagem, utiliza-se informações estruturais das proteínas depositadas no PDB, nas quais utiliza-se somente as coordenadas tridimensionais dos átomos de C_α para representar cada aminoácido da cadeia estrutural da proteína.

II. Problema da Geometria de Distâncias Moleculares (PGDM)

Define-se o Problema da Geometria de Distâncias Moleculares (PGDM) como sendo: encontre as coordenadas tridimensionais $x_1, x_2, \ldots, x_n$ dos átomos de uma proteína de modo que

$$\|x_i - x_j\| = d_{ij}, \text{ para } (i,j) \in S, \tag{1}$$

onde S é um subconjunto de pares de átomos cujas distâncias são conhecidas e $\|\cdot\|$ é a norma Euclidiana.

Para resolver o PGDM, existem duas principais abordagens: na primeira, pesquisadores formulam um problema de otimização não-linear e empregam métodos contínuos para obterem uma solução aproximada; na segunda, o problema é resolvido utilizando técnicas de otimização discreta e teoria de grafos. Neste trabalho, em específico, utiliza-se a abordagem contínua, que, dentre ambas, é a mais clássica.

O objetivo deste estudo consiste em reconstruir a estrutura tridimensional de uma proteína sendo conhecidas somente as distâncias intra-átomos. Para isso, resgata-se os dados estruturais das proteínas do *Protein Data Bank* (PDB) e calcula-se todas as distâncias entre

os átomos de Carbono Alfa (C_α), visto que as estruturas podem ser representadas, mais facilmente e sem grande perda de características, pelas coordenadas dos átomos de C_α de cada aminoácido.

Para resolver o PGDM, pode-se formular o seguinte problema de otimização não-linear:

$$\begin{aligned} &\min \sum \left(\|x_i - x_j\|^2 - d_{ij}^2\right)^2 \\ &\text{sujeito a } x_i \in \mathbb{R}^3, i = 1, 2, \ldots, n. \end{aligned} \tag{2}$$

Os métodos utilizados para resolver numericamente o problema (1) são: BFGS e Gauss-Newton. O primeiro é caracterizado por ser um método quase-Newton de otimização irrestrita, por aproximar a matriz Hessiana da função objetivo, contornando o alto gasto computacional do método de Newton clássico no cálculo da direção de busca. O segundo é um dos mais clássicos métodos de resolução de sistemas não-lineares e problema de mínimos quadrados não-lineares, cuja direção de busca é determinada por um sistema linear envolvendo a matriz Jacobiana da função objetivo.

III. Resultados Computacionais

Para os experimentos numéricos implementamos os métodos BFGS e Gauss-Newton, bem como os algoritmos auxiliares, no *software* MATLAB[1]. Além disso retiramos oito proteínas do *Protein Data Bank* e tomamos somente as coordenadas dos átomos de carbono alfa (C_α) presentes em cada estrutura. A sigla de cada proteína e o número de átomos C_α estão presentes na Tabela I.

Tabela I: Proteínas utilizadas nos experimentos.

Proteína	n_{C_α}
1A9W	572
1AQR	40
1BQX	77
1BSA	321
1F39	202
1FS3	124
1JK2	90
1MBN	153

Nestes experimentos, foram selecionados todos os átomos de C_α das oito proteínas da Tabela I e calculados os conjuntos completos de distâncias exatas. Repetiu-se cada problema 20 vezes, sempre partindo de um ponto inicial gerado aleatoriamente e os melhores resultados obtidos são apresentados nas Tabelas II e III, cujas colunas obedecem a seguinte legenda: Proteína é a sigla da proteína retirada do PDB, $nvar$ é o número de variáveis do problema, $iter$ é o total de iterações, $f(x^*)$ é o valor da função na solução e $t(s)$ é o tempo de execução medido em segundos. Em todos os experimentos o número máximo de iterações foi 400 e a precisão de 10^{-4}.

[1] https://www.mathworks.com/products/matlab.html

Tabela II: Resultado dos Testes Numéricos - BFGS

Problema		BFGS		
Proteína	$nvar$	$iter$	$f(x^*)$	$t(s)$
1A9W	1716	53	$5.9390e-03$	39.7073
1AQR	120	45	$2.2945e-05$	0.1629
1BQX	231	51	$1.2723e-05$	0.5621
1BSA	963	91	$2.0218e-03$	20.1517
1F39	606	81	$2.3484e-03$	6.6483
1FS3	372	57	$8.8267e-04$	1.6664
1JK2	270	77	$9.0695e-04$	1.1818
1MBN	459	49	$6.0768e-04$	2.2029

Tabela III: Resultado dos Testes Numéricos - Gauss-Newton

Problema		Gauss-Newton		
Proteína	$nvar$	$iter$	$f(x^*)$	$t(s)$
1A9W	1716	26	$1.4164e-20$	205.5401
1AQR	120	30	$1.3944e-24$	0.0829
1BQX	231	24	$1.1852e-23$	0.3156
1BSA	963	26	$6.0874e-21$	26.2482
1F39	606	20	$9.7204e-22$	4.2060
1FS3	372	15	$8.7508e-23$	0.6524
1JK2	270	23	$1.1712e-22$	0.3994
1MBN	459	31	$1.7444e-22$	2.4517

IV. Conclusão e Perspectivas Futuras

Os resultados obtidos neste trabalho evidenciam o melhor desempenho, em partes, do método de Gauss-Newton, sobretudo no baixo número de iterações e a qualidade da solução, com um valor muito próximo de zero. Por outro lado, o método BFGS também conseguiu resolver todos os problemas e com menor tempo computacional gasto, principalmente nos problemas com um grande número de variáveis.

Em relação aos resultados numéricos obtidos, principalmente em relação ao tempo computacional gasto pelo método de Gauss-Newton, tem-se como perspectivas futuras realizar modificações no cálculo da direção de busca, no qual foi utilizado a função "/" do MATLAB (por *default* resolve o sistema linear por fatoração LU). Serão realizados estudos para verificar outras opções de abordar esse problema. Além disso, para trabalhos futuros, será analisado o comportamento dos métodos abordados neste trabalho na resolução do PGDM com um conjunto incompleto de distâncias.

Referências

[1] R. S. Lima, Reconstrução e classificação de estruturas espaciais via otimização contínua: ênfase em proteínas, Tese de Doutorado em Matemática Aplicada, Unicamp, (2012).

[2] A. Mucherino, C. Lavor, L. Liberti, and N. Maculan. *Distance Geometry: Theory, Methods, and Applications*. Springer, New York, 2013.

[3] A. A. Ribeiro, e E. W. Karas. *Otimização Contínua: Aspectos Teóricos e Computacionais*. Cengage, São Paulo, 2013.