Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Representação de Proteínas usando Geometria Hiperbólica

João Alexandre R. A. M. Souza[1], Henrique Vitório[2]
DMAT/UFPE, Recife, PE
Carlile Lavor[3]
Unicamp, Campinas, SP

A estrutura tridimensional de uma proteína é um dos principais fatores que determinam a sua interação com outras moléculas, sejam essas carboidratos, lipídios, outras proteínas, ou mais.

A fim de estudar a distribuição das estruturas de proteínas conhecidas, vemos útil fixar uma medida de similaridade entre cada par de proteínas, uma vez que tal medida nos permitirá usar técnicas de **redução de dimensionalidade** para analisar relações num conjunto de proteínas. Tal redução de dimensionalidade é indispensável, visto que a estrutura de cada proteína é pensada como um ponto em $\mathbb{R}^{3N}$, onde N é o número de átomos desta, que pode vir a ser grande.

Em primeiro momento, fixamos a **Root Mean Square Deviation** (**RMSD**) [2] como uma medida de similaridade, ou, mais precisamente, como uma medida de **dissimilaridade**. Para obter uma boa visualização das relações entre proteínas num conjunto, procuramos por representações em dimensões baixas para o conjunto: cada proteína será representada por um ponto num espaço de duas ou três dimensões. Tais representações, ditas **embeddings**, são obtidas por métodos computacionais.

É possível **inferir funções de uma proteína a partir de quais outras proteínas possuem representações próximas a ela** [1], o que motiva um processo de classificação: dada a estrutura de uma proteína específica, é possível determinar sua função a partir de qual vizinhança de proteínas aquela se encontra.

O mapeamento de proteínas a um espaço de dimensão menor para avaliar relações entre elas já foi realizado com sucesso no espaço euclidiano $\mathbb{R}^n$ [1], mas, neste espaço, o fenômeno de distorção inerente a técnicas de redução de dimensionalidade se torna evidente, e a representação obtida pode não corresponder bem às relações reais. Motivados pelo fato que o **espaço hiperbólico** $\mathbb{H}^n$ possui propriedades distintas que permitem **representações de grafos com estruturas hierárquicas com distorção arbitrariamente baixa** [4], usamos as técnicas de obtenção de embeddings hiperbólicas apresentadas em [3] para obter representações de conjuntos de proteínas no plano hiperbólico.

Denote por $S = [s_{ij}]$ a matriz quadrada das dissimilaridades s_{ij} entre proteínas i e j, e denote por $D = [d_{ij}]$ a matriz quadrada de distâncias entre as representações obtidas para essas proteínas. A quantidade

$$\text{Stress}(D) = \frac{\|S - D\|_F^2}{\|S\|_F^2} = \frac{\sum_{ij}(s_{ij} - d_{ij})^2}{\sum_{i,j} s_{ij}^2}, \tag{1}$$

dita estresse normalizado, é uma medida do quanto as distâncias da representação concordam com as dissimilaridades de RMSD dadas [1]. Quanto menor tal quantia, melhor a representação obtida reflete as dissimilaridades originais. Em experimentos iniciais para um conjunto de proteínas fixado, observamos que tal quantia foi, consistentemente, **menor no espaço hiperbólico do**

[1]joaoramatta@hotmail.com
[2]henrique.vitori@ufpe.br
[3]clavor@unicamp.br

quê no espaço euclidiano. Tal indicativo sugere que o espaço hiperbólico é um ambiente mais adequado para o tipo de mapeamento a dimensões menores em questão.

O espaço hiperbólico é um exemplo de espaço de curvatura constante e negativa, onde áreas e volumes de discos crescem **exponencialmente** com o raio do disco, em contraste com o crescimento **polinomial** presente no espaço euclidiano. Como consequência, o espaço hiperbólico se vê mais "espaçoso" e é mais adequado como espaço ambiente para representação geométrica de dados, ainda mais quando esses dados estão distribuídos de maneira hierárquica.

Exploramos o uso do espaço hiperbólico na análise de estruturas de proteínas. A estrutura tridimensional de uma proteína é crucial para determinar suas interações moleculares. Para analisar a distribuição das estruturas de proteínas conhecidas, utilizamos a Root Mean Square Deviation (RMSD) como uma medida de dissimilaridade e empregamos técnicas de redução de dimensionalidade ao problema.

Através de métodos computacionais, buscamos representações de baixa dimensão para o conjunto de proteínas, permitindo a inferência de funções das proteínas com base em suas proximidades a outras proteínas no espaço de representação.

Embora o mapeamento de proteínas para um espaço de dimensão menor tenha sido realizado com sucesso no espaço euclidiano, a distorção inerente às técnicas de redução de dimensionalidade pode afetar a qualidade do resultado. Exploramos o uso do espaço hiperbólico, o qual permite representações de grafos com estruturas hierárquicas com distorção arbitrariamente baixa, como uma escolha mais adaptada para um espaço de representação para o nosso mapa.

Nossos experimentos iniciais indicam que a qualidade do espaço hiperbólico como espaço de representação é, de fato, melhor do que a qualidade do espaço euclidiano.

O espaço hiperbólico, um exemplo de espaço de curvatura constante e negativa, apresenta crescimento exponencial de áreas e volumes de discos com o raio, tornando-o mais "espaçoso" e adequado para a representação geométrica de dados, especialmente quando esses dados estão distribuídos de maneira hierárquica.

Agradecimentos

Agradecemos à CAPES por fomentar o doutorado que está produzindo o presente trabalho.

Referências

[1] J. Hou, S.-R. Jun, C. Zhang e S.-H. Kim. "Global mapping of the protein structure space and application in structure-based inference of protein function". Em: **Proceedings of the National Academy of Sciences** 102.10 (2005), pp. 3651–3656. DOI: 10.1073/pnas.0409772102. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.0409772102. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0409772102.

[2] I. Kufareva e R. Abagyan. "Methods of Protein Structure Comparison". Em: **Homology Modeling: Methods and Protocols**. Ed. por Andrew J. W. Orry e Ruben Abagyan. Totowa, NJ: Humana Press, 2012, pp. 231–257. ISBN: 978-1-61779-588-6. DOI: 10.1007/978-1-61779-588-6_10. URL: https://doi.org/10.1007/978-1-61779-588-6_10.

[3] M. Nickel e D. Kiela. "Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry". Em: **International Conference on Machine Learning**. 2018.

[4] C. De Sa, A. Gu, C. Ré e F. Sala. "Representation Tradeoffs for Hyperbolic Embeddings". Em: **Proceedings of the 35th International Conference on Machine Learning**. Ed. por Jennifer Dy e Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, out. de 2018, pp. 4460–4469. URL: https://proceedings.mlr.press/v80/sala18a.html.