Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 10, n. 1, 2023.

Trabalho apresentado no XLII CNMAC, Universidade Federal de Mato Grosso do Sul - Bonito - MS, 2023

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Geometria Hiperbólica Aplicada à Clusterização de Proteínas

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

Proteínas realizam suas funções ao se ligarem a outras moléculas como carboidratos, lipídios, outras proteínas, etc. A interação ou não de uma proteína com outra molécula é principalmente determinada pela estrutura 3D da proteína, o que motiva a **classificação e mapeamento do espaço universo de proteínas de acordo com suas similaridades**.

A fim de explorar o espaço universo de proteínas, podemos tomar a **Root Mean Square Deviation** como uma métrica de quão similar são as estruturas 3D de duas proteínas [1]. O cenário obtido é de um grafo completo não-direcionado, onde seus vértices representam proteínas e o peso nas arestas é dado pela medida de similaridade entre elas. Em tal âmbito, podemos realizar uma imersão do espaço que desejamos estudar (o espaço de proteínas) a um espaço de similaridade, de forma a obter uma visualização geral do espaço de proteínas, onde devemos observar um fenômeno de **clusterização** de vértices com altos graus de similaridade [2].

O fenômeno de clusterização pode vir a auxiliar uma análise comparativa de função e forma de proteínas com relação às suas estruturas 3D, isto é, é possível **inferir funções de uma proteína a partir de quais outras se encontram próximas a ela no espaço de similaridade** [3], o que também abre caminho para revelar princípios que determinam as estruturas das proteínas.

Voltando nosso foco ao fenômeno de clusterização mencionado, podemos deduzir haver uma **hierarquia entre as moléculas** de acordo com sua classificação de similaridade. A partir desta observação, a inspiração a usar o espaço hiperbólico se torna aparente, uma vez que o espaço hiperbólico possui propriedades distintas que permitem **representações de grafos com estruturas hierárquicas com distorção arbitrariamente baixa** [4]. O espaço hiperbólico é um exemplo de espaço de curvatura constante negativa, o que permite, por exemplo, que áreas de discos e volumes de bolas cresçam **exponencialmente** com o raio ao invés de polinomialmente como ocorre no espaço euclidiano. Como consequência, o espaço hiperbólico é mais "espaçoso" e está mais apto a abranger mergulhos geométricos de dados, especialmente quando estes possuem algum tipo de estrutura hierárquica. Tratando-se do espaço hiperbólico, existem diversos modelos distintos, todos equivalentes, e podemos escolher aquele que sirva mais à situação. Dependendo da situação, o modelo do disco de Poincaré pode ser o modelo favorecido, uma vez que oferece uma boa visualização no plano.

No espaço hiperbólico, o fenômeno de clusterização é exacerbado e este pode vir a revelar mais relações através da inferência de proximidade, como é observado na figura 1, onde cada língua é representada por um ponto no espaço hiperbólico e essas são comparadas de acordo com o número de palavras em comum e de mesmo sentido que compartilham.

No presente trabalho, realizado como início à pesquisa de doutorado, pretendemos fazer uso do espaço hiperbólico como ambiente para comparação de similaridades entre estruturas moleculares,

[1]joao.matta@ufpe.br
[2]henrique.vitori@ufpe.br
[3]clavor@unicamp.br

com intento de revelar clusterizações e hierarquias que permitam um mapeamento global do espaço das moléculas de proteínas.

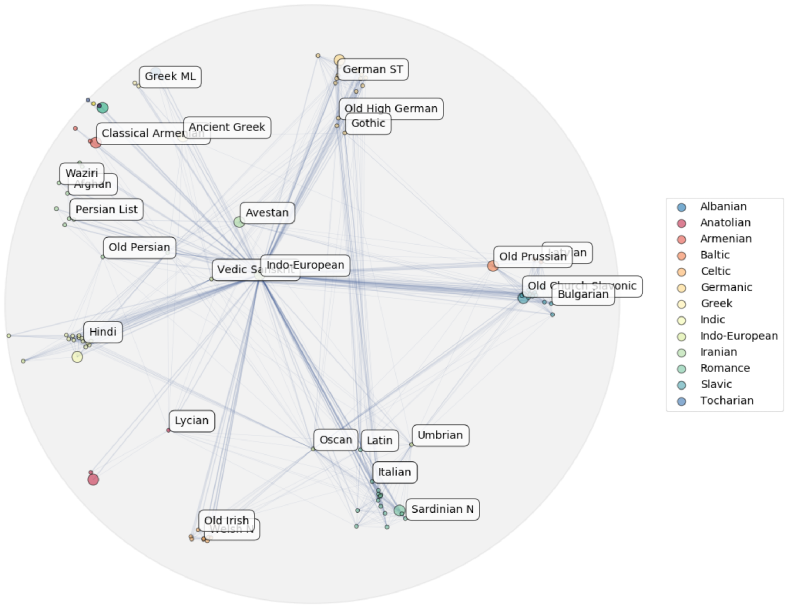

Figura 1: Figura retirada do artigo [2], demonstrando, no plano hiperbólico, um efeito de clusterização de línguas de acordo com o banco de dados IELex.

Agradecimentos

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

Referências

[1] Irina Kufareva e Ruben 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.

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

[3] Jingtong Hou, Se-Ran Jun, Chao Zhang e Sung-Hou 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.

[4] Frederic Sala, Chris De Sa, Albert Gu e Christopher Re. "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.