Trabalho apresentado no CNMAC, Gramado - RS, 2016.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Classificação de Proteínas Através de Homologia Persistente

Sabrina S. Calcina[1]
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP
Marcio Gameiro [2]
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP

1 Introdução

O objetivo deste trabalho é utilizar *homologia persistente* para classificar um conjunto de proteínas. Homologia persistente [2] é uma ferramenta que permite calcular as propriedades topológicas de um objeto e sua robustez com relação às mudanças nos parâmetros. Mais precisamente, homologia persistente conta o número de componentes conexas e buracos de várias dimensões, e mostra o quanto eles persistem. Considere a sequência crescente de complexos simpliciais, chamada *filtração*, $K^0 \subset K^1 \subset \cdots \subset K^5$ da Figura 1. Para

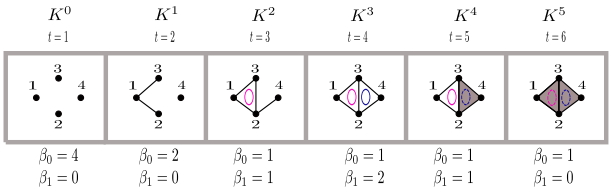

Figura 1: Sequência crescente de complexos simpliciais.

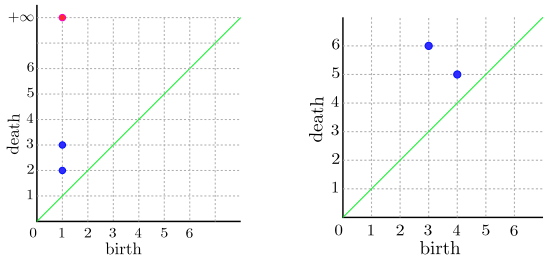

Figura 2: Diagramas de persistencia da filtração da Figura 1 correspondente as componentes conexas β_0 (esquerda) e aos ciclos β_1 (direita).

cada complexo K^i nesta filtração, o número de componentes conexas β_0 e o número de ciclos β_1 é indicado na Figura 1. Homologia persistente, representada pelos *diagramas*

[1]sabrinas@icmc.usp.br
[2]gameiro@icmc.usp.br

de persistencia da Figura 2, nos diz por quanto tempo cada uma destas propriedades topológicas persistem. O ponto $(3, 6)$ no diagrama correspondente a β_1, por exemplo, nos diz que um ciclo foi criado no instante $t = 3$ e destruído no instante $t = 6$. O ponto $(1, +\infty)$ no diagrama correspondente a β_0 indica que uma das componentes conexas que foram criadas no instante $t = 1$ nunca desapareceu.

Assim, este trabalho consiste em calcular a homologia persistente de um conjunto de proteínas e utilizar esta informação topológica para classificar estas proteínas.

2 Metodologia e resultados preliminares

Dado um conjunto de proteínas, calculamos as filtrações de alpha complexos com pesos (`www.cgal.org`) e calculamos a homologia persistente destas filtrações [2]. Então, utilizamos os diagramas de persistencia correspondentes aos ciclos (túneis) β_1 e as cavidades β_2 para calcular a matriz das distâncias Wasserstein [1] entre estes diagramas. Finalmente, utilizamos esta matriz de distâncias para classificar as proteínas de acordo com suas propriedades topológicas.

Como um teste inicial para o método proposto, tomamos um conjunto de 218 proteínas extraídas do banco de dados PDB (`www.rcsb.org`). Este conjunto foi escolhido a fim de formar dois grupos distintos, um com 166 proteínas e o outro com 52. Aplicando o método descrito acima obtemos as matrizes de distâncias apresentadas na Figura 3. Claramente, obtemos dois grupos distintos de proteínas, e estes grupos correspondem precisamente aos dois grupos com 166 e 52 proteínas, conforme esperado.

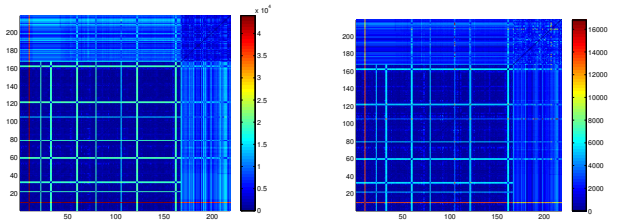

Figura 3: Matriz de distâncias Wasserstein correspondente aos túneis β_1 (esquerda) e as cavidades β_2 (direita).

3 Conclusões

Baseado no teste preliminar apresentado, o método parece promissor, porém precisa ser testado em conjuntos maiores para validar sua eficácia. Além disso, pretendemos utilizar técnicas de agrupamento baseadas em matrizes de distâncias para obter os agrupamentos dos dados de acordo com sua topologia.

Referências

[1] D. Cohen, H. Edelsbrunner, J. Harer, Stability of Persistence Diagrams, *Discrete and Computational Geometry*, v. 1, (2007). DOI: 10.1007/s00454-006-1276-5.

[2] H. Edelsbrunner, *A Short Course in Computational Geometry and Topology*. Springer, (2014).