How to make bioinformatics FAIR: the dark side of overpromising and overselling

Michał Burdukiewicz
Autonomous University of Barcelona
Medical University of Białystok
National Institute of Public Health

7 June 2022

Abstract

Some bioinformatics tools have a profound and evident influence on future research and industry applications. For example, the BLAST algorithm for identifying similarities between biological sequences was cited in almost 5000 patent families. However, due to the nature of bioinformatics (relatively low amount of resources needed to conduct the research in the field), we are now dealing with the influx of low-quality, oversold tools that wrongly claim to solve problems that they are addressing.

During the presentation, I will showcase two case studies: prediction of antimicrobial peptides and amyloid-amyloid interactions.

Antimicrobial peptides (AMPs) are short cationic peptides with antimicrobial activity. Based on our results, we show that existing models for antimicrobial peptide prediction overpromise their factual accuracy. Next, we propose AMPBenchmark, a unified framework for fair benchmarking of such tools.

AmyloGraph is a database of interactions between amyloids, self-aggregating proteins which occur, among others, in neurodegenerative disorders. Due to the rigorous data curation procedure, we can prove that the current experimental data is unsuitable for building a robust predictor of amyloid-amyloid interactions.

About the presenter

Michał Burdukiewicz is a bioinformatician affiliated with the Autonomous University of Barcelona, the Medical University of Białystok, and the National Institute of Public Health. His scientific interests involve machine learning applications in the analysis of protein sequences, especially sampling of negative data. In his free time, Michał Burdukiewicz popularizes data science by organizing *Why R?* conferences and related events. Moreover, he coordinates an award for Women in Data Science.