BRIGHAM AND WOMEN'S HOSPITAL
HARVARD MEDICAL SCHOOL

181 Longwood Avenue
Boston, Massachusetts 02115-5804

Department of Medicine
Channing Division of Network Medicine

Channing Microbiome Seminar

March 8 (Friday), 2024, 9AM (ET)

MCP 5th-floor large conference room & Zoom:
https://us02web.zoom.us/j/81070959105?pwd=RFJNd3dSZmR6dXJZNjJiYVVzQ3NEQT09
Meeting ID: 810 7095 9105; Passcode: 984617

Yiyan Yang, PhD

National Library of Medicine, National Institutes of Health

Large-scale Genomic Survey with Deep Learning-based Method Reveals Strain-Level Phage Specificity Determinants

Background: Phage therapy, re-emerging as a promising approach to counter antimicrobial-resistant infections, relies on a comprehensive understanding of the specificity of individual phages. Yet the significant diversity within phage populations presents a considerable challenge. Currently, there's a notable lack of tools designed for large-scale characterization of phage receptor-binding proteins, which are crucial in determining the phage host range. Results: In this study, we present SpikeHunter, a deep learning method based on the ESM-2 protein language model. With SpikeHunter, we identified 231,965 diverse phage-encoded tailspike proteins, a crucial determinant of phage specificity that targets bacterial polysaccharide receptors, across 787,566 bacterial genomes from five virulent, antibiotic-resistant pathogens. Notably, 86.60% (143,200) of these proteins exhibited strong associations with specific bacterial polysaccharides. We discovered that phages with identical tailspike proteins can infect different bacterial species with similar polysaccharide receptors, underscoring the pivotal role of tailspike proteins in determining host range. The specificity is mainly attributed to the protein's C-terminal domain, which strictly correlates with host specificity during domain-swapping in tailspike proteins. Importantly, our dataset-driven predictions of phage-host specificity closely match the phage-host pairs observed in real-world phage therapy cases we studied. Conclusions: Our research offers a comprehensive resource that encompasses both a methodology and a database from an extensive genomics survey. This significantly improves our understanding of phage specificity determinants at the strain level, while also providing a robust framework for selecting phages in therapeutic contexts. Additionally, this approach can be extended to identify bacteria-phage interactions at more precise taxonomic levels within the microbiome, paving the way for future in-situ microbiome engineering.

Bio: I identify myself as a bioinformatician dedicated to utilizing computational techniques to address challenges in microbiology. Currently, I am a postdoctoral researcher at the National Library of Medicine, National Institutes of Health. In 2019, I received my doctoral degree from a bioinformatics graduate program at Tongji University in Shanghai, China. My PhD thesis focused on exploring host-microbe and environment-microbe interactions, particularly from an evolutionary perspective. I commenced my postdoctoral training in 2020 at the National Library of Medicine (NLM), where I not only continued my existing research focus but also expanded it to include large-scale genomic analysis for studying coronaviruses and bacteria. My research contributions span areas including bacteria-phage interactions, microbial genotype-phenotype correlations, and coronavirus evolution, which has been acknowledged in publications such as Molecular Biology and Evolution, GigaScience, mSystems, and Bioinformatics. In summary, my primary research interest lies at the intersection of computer science and biology, specifically focusing on microbiology and virology. My future goal is to contribute to the understanding of the role microbes play in diagnosing, treating, and progressing human diseases.