Synthesis and Motif Discovery of Novel Antibacterial Molecules and Orally Active Drugs via Sequence-Based Machine Learning

Wang, Tony (School: Amador Valley High School)

Each year, over two million people globally are afflicted by potentially life-threatening antibiotic-resistant infections, and many scientists believe that antibiotic-resistant superbugs may cause the next pandemic. Not only that, the current oral drugs being developed are failing to pass clinical trials or meet even the most basic pharmacological criteria. Thus, the objective of this research was to develop a machine learning approach to computationally generate novel antibacterial and orally active molecules. Training data was gathered from the ChEMBL repository, which was split into antibacterial protein inhibitors and medicinal drug-like molecules. Then, Gated Recurrent Unit (GRU) neural network models were trained on the datasets and used to generate novel sequences, treating chemical sequences as words in a language. The generated sequences were evaluated based on machine learning classifiers and pharmacological criteria, and the GRU is retrained on the top molecules from the previous iteration. Through this method, thousands of novel molecules were generated. Then, various sequence analyses were run on the generated molecules to identify recurring motifs by generating novel input sequences, ranking, and clustering. Multiple recurring sequence motifs were identified, which may be valuable in future studies, and the motif robustness was tested through correlation analyses and synthetic data simulations. Last, the versatility of the methods were demonstrated through applications to protein amino acid sequences. Overall, the versatile methods developed in this research can potentially be reused to design molecules for a wide range of purposes, such as in environmental science, and ultimately lead to faster, more effective molecular design.