Authors: J.Santerre, F. Xia, S.Boisvert, R.Stevens, Argonne National Laboratory

Title: Towards antimicrobial resistance classification as a service

Abstract: Recent advances in DNA sequencing throughput and accuracy accompanied by plummeting per-base cost is making sequence-based applications more amenable. While a plethora of web platforms are available for convenient and push-button analysis (e.g. Galaxy, DNAnexus, One Codex, etc), tools that can decipher patterns from labeled data are not yet available to biologist as an easy-to-use web platform. Here we present a web platform that supports large-scale analysis of sequence with corresponding antimicrobial resistance data.

We are pursuing techniques and developing tools that enable statistical inference on reads directly obtained from DNA sequencers. In our first work we use Random Forests (RF), a naively parallelizable and established Machine Learning algorithm, to produce classifiers that label out-of-sample strains as resistant (RES) or susceptible (SUS) after training on a population of known strains. While our results are promising, (95% accuracy, correct identification of known genes responsible for resistance), there are many more techniques that can be applied to biological problems. We believe one central outcome of cloud computing in biology will be the full integration of such tools and we hope to help usher in that utilization.

The framework for a typical study related to RES and SUS populations is as follows: a biologist repeatedly exposes an antibacterial agent to a population, thereby forcing the development of genetic mutations that confer resistance. When resistance has been induced, the new strain is sequenced and compared to the original population, also called the wild type. In such controlled environments, the set of significant mutations that distinguishes the difference between the RES and SUS groups is considered the set responsible for resistance. Using K-mers as features, we train the RF to determine the optimal set of K-mers for classification of a novel strain as RES or SUS. Additionally, RF provides a quantification of the importance of each K-mer, which allows us to identify the location of key mutations.

In our first analysis we show that RF, performed on both contigs and reads data alone, is able to identify with high accuracy the difference between SUS and RES populations of *S. pneumoniae* and *Mycobacterium tuberculosis* with accuracies on different datasets as low as 80% (100 samples) and as high as 95% (3000 samples). Beyond classification accuracy, we cluster the significant K-mers by gene function from a reference genome and note that the most important features have existing literature indicating involvement in resistance and susceptibility. In short, RF is appears to be robust, and despite lower accuracy on fewer strains (100 vs. 3000) it still is able to correctly identify genes known to be involved in antibacterial resistance.

While we have chosen to train on K-mers in our first work, in theory other aspects of the individual strains could be used (ie. survival rates, GC content, virulence). Other algorithms may provide even more potential for analyzing stochastic processes, for considering the network of mutations conferring resistance, or for ranking the importance of the mutations.

This work was presented at GLBio 2015(International Society for Computational Biology) where it received a best poster award. It will be presented as a 15 minute talk at the ALCoB 2015 in the beginning of August and is in draft as a case study for ISCB journal.