Automatic Identification of Resistance-Associated Mutations Using Techniques from Human Language Technologies

Betty Y. Cheng[1], Jaime G. Carbonell[1]

[1] Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pennsylvania, USA

OBJECTIVES: As more antiretroviral drugs and drug resistance data become available, the need for a statistical system to predict HIV phenotype from genotype intensifies. However, most statistical prediction systems use features selected by human experts which lead to the same bottleneck found in maintaining rule-based systems. Here, we developed a method to automatically identify resistance-associated mutations that can perform as well as features selected by human experts in drug resistance prediction.

METHODS: We adapted an approach used to identify keywords in text documents to automatically generate a feature set of resistance-associated mutations for HIV phenotype prediction. Viewing each HIV genotype as a text document in a language without word boundaries, n-grams of varying lengths were extracted at each reading frame as well as position-specific unigrams, and their counts were tabulated. For each n-gram count, we derived 20 binary features by considering whether the n-gram occurred at least *j* times, and computed the chi-square statistic of each binary feature to measure its ability to discriminate between susceptible, low-resistant and high-resistant HIV genotypes. N-grams were then ranked by the highest chi-square statistic from their derived binary features. The top *p* features were used with standard machine learning methods to predict phenotype where *p* was optimized for each antiretroviral drug.

RESULTS: Using the same classifier (Decision Tree) and dataset as a previous comparative study on feature sets for phenotype prediction (Rhee *et al.*, PNAS 2006), our position-independent features, position-specific features and their mixture yielded accuracies of 0.787, 0.784, and 0.788 respectively averaged across all drugs, which were comparable to the reported accuracies of 0.775 from mutations selected by human experts and 0.784 from mutations trained using special treatment history data. Moreover, when used with random forest classifier, our feature sets yielded improved accuracies of 0.809-0.812. Further analysis showed the human expert selected mutations to closely overlap with our chi-square selected features.

CONCLUSIONS: Contrary to a previous study, we developed an automatic method to identify resistance-associated mutations for phenotype prediction that can match the performance of human experts without special treatment history data. This method can remove the human bottleneck in statistical prediction systems relying on human-selected mutations.