Identifying hosts of viral families: A Machine Learning approach

Anil Raj[1], Mike Dewar[1], Raul Rabadan[2,3], Chris H. Wiggins[1,3]
[1]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY
[2]Department of Biomedical Informatics, Columbia University, New York, NY
[3]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY

Motivated by the framework in Leslie et al. [**?**], we represent each viral protein sequence in a feature space spanned by the set of all possible subsequences of amino acids of a fixed length k. Each protein is then represented by counts of (k,m)-patterns generated by all k-length subsequences in its sequence, where the (k,m) pattern generated by a string of length k is the set of all k-length strings differing from it by at most m mismatches. Non-zero values for m allow us to better capture rapidly mutating, yet conserved, viral subsequences (or genomic regions) that are informative of the host of the virus. Each protein is also assigned one of 3 labels — invertebrate, plant or vertebrate — depending on the host of the virus from which the protein was extracted.

Armed with this representation, we use multi-class Adaboost to learn sparse, interpretable models to predict the host of a viral protein, from which the host of the virus can then be discerned. The model learned in this problem is an Alternating Decision Tree that outputs a real-valued vector prediction for each input protein — the dimensionality of the prediction equals the number of classes. The viral protein is then assigned to the class with the largest predicted output.

In this work, we attempted to build a predictive and interpretable model to classify viruses belonging to the family *Picornaviridae* — a family of viruses that contain a single stranded, positive sense RNA less than 10 kilobases in length. The accuracy of the alternating decision tree model, at each round of boosting, was evaluated using a multi-class ROC score. At each point on the ROC curve, a protein is considered to be classified correctly if the real-valued output for the true class of that protein is greater than the threshold value corresponding to that point on the ROC curve. We observe that the *smoothing* effect introduced by using the mis-match feature space allows for improved prediction accuracy for larger values of m. To visualize the predictive k-mer subsequences selected by Adaboost, we mark the k-mers on all viral protein sequences grouped by their label, with their lengths scaled to $[0, 1]$ - this helps elucidate regions of protein sequences that are conserved in viruses of a specific host. Encouraged by the success of this approach to classifying picornaviruses, future work will involve extending

this approach to more difficult to classify virus families.