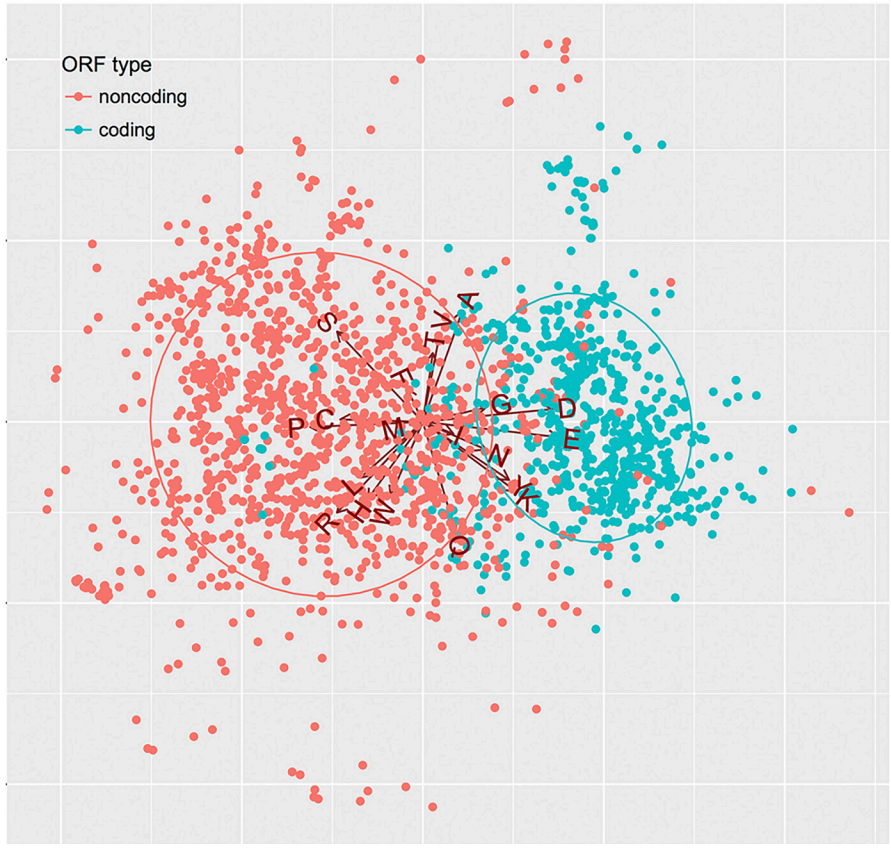
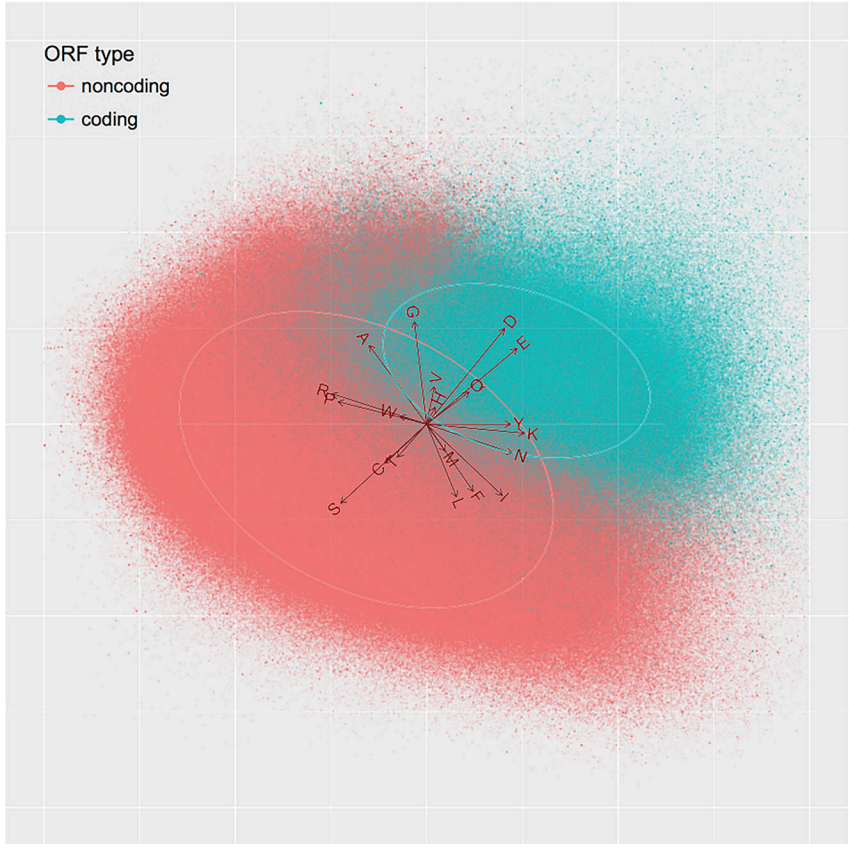
Using Machine Learning to Find Open Reading Frames that Code for a Protein

Given that an open reading frame is simply a stretch of DNA between stop codons, which account for 3 out of the 64 codons, we are left with a large pool of potential protein coding ORFs. There are many different methods to differentiate ORFs that code for a protein from those that occur by chance. The most prevalent approach is using shared homology of candidate ORFs to a database of known ORFs. This method presents many pitfalls, most notably the catch that you only find genes that you have seen before. Instead we have investigated using the amino-acid frequency of open reading frames to determine which code for a protein, and which are spurious non-coding. We utilize the fact that proteins have a conserved amino-acid composition that is different from the amino-acid composition in non-coding open reading frames. There are two classes of non-coding open reading frames: those that occur within intergenic DNA stretches and those that occur within the nucleotides of a gene but in an overlapping frame (Figure 1). Intergenic ORFs have an amino-acid frequency independent of coding frequency, while overlapping-alternate-frame ORFs have frequencies dependent on the amino-acid frequency of the overlapped coding frame. In relation to this coding frame (denoted as position +0), there are two same strand frames (+1 and +2) and three opposite strand frames (-0, -1, and -2). Thus we would expect there to be one conserved amino acid frequency for the coding frame, and five spurious frequencies each corresponding to the alternate frames. In addition there will be one or more frequencies for intergenic regions that would only depend on the GC content and how often you would expect to see the various amino-acids by chance. Plotting these coding and noncoding ORFs in 20 dimensional space allows us to visualize the discrete clusters. We can then use machine learning to differentiate the clusters, so that we can then pick out the coding ORFs and discard the rest.

Katelyn McNair *and Robert Edwards*

This research is supported by the Computational Science Research Center (CSRC) at San Diego State University

Plot of amino-acid frequency of Lambda virus ORFs



Plot of amino-acid frequency of all ORFs