In Silico Predictive Modelling of CRISPR/Cas9 guide efficiency

The CRISPR/Cas9 system provides state-of-the art genome editing capabilities. However, several facets of this system are under investigation for further characterization and optimization. One in particular is the choice of guide RNA that directs Cas9 to target DNA--given that one would like to target the protein-coding region of a gene, hundreds of guides satisfy the constraints of the CRISPR/Cas9 Protospacer Adjacent Motif sequence. However, only some of these guides efficiently target DNA to generate gene knockouts. One could laboriously and systematically enumerate all possible guides for all possible genes and thereby derive a dictionary of efficient guides, however, such a process would be costly, time-consuming, and ultimately not practically feasible. Instead, one can (1) enumerate all possible guides over each of some smaller set of genes, and then test these experimentally by measuring the knockout capabilities of each guide, (2) thereby assemble a training data set with which one can “learn”, by way of predictive machine learning models, which guides tend to perform well and which do not, (3) use this learned model to generalize the guide efficiency for genes not in the training data set. In particular, by deriving a large set of possible predictive features consisting of both guide and gene characteristics, one can elicit those characteristics that define guide-gene pairs in an abstract manner, enabling generalizing beyond those specific guides and genes, and in particular, for genes which we have never attempted to knock out and therefore have no experimental evidence. Based on such a set of experiments, we present a state-of-the art predictive approach to modeling which RNA guides will effectively perform a gene knockout by way of the CRISPR/Cas9 system. We demonstrate which features are critical for prediction (e.g., nucleotide identity), which are helpful (e.g., thermodynamics), and which are redundant (e.g., microhomology). Finally, we combine our insights of useful features with exploration of different model classes, settling on one model which performs best. Finally, we elucidate which measures should be used for evaluating these models in such a context.