Spring 2018

ceFinder: Machine Learning Based Prediction of Novel Competing Endogenous RNAs

Teng Gao
Washington University in St. Louis

ceFinder: Machine Learning Based Prediction of Novel Competing Endogenous RNAs

Teng Gao

Mentors: Ha X. Dang and Christopher A. Maher

Despite the growing importance and biological relevance of long noncoding RNAs (lncRNAs) in disease biology, many challenges remain in dissecting their regulatory mechanisms. Competitive endogenous RNA (ceRNA) hypothesis attracted much attention as one of the major regulatory mechanisms of long non-coding RNAs in cancers. A ceRNA interaction triplet is characterized by the three-way interaction between a target mRNA and a long non-coding RNA (referred to as a ceRNA) that competes to bind with one or more microRNAs (miRNAs). Because currently available experimental methods to validate novel ceRNAs are expensive and low-throughput, various computational approaches have been developed to predict novel ceRNAs. However, the absence of a gold-standard dataset resulted in the lack of rigorous performance validation of the current in-silico prediction algorithms. Moreover, individual algorithm often performed poorly on new data. To address these problems, the present study aims to construct an experimentally validated ceRNA dataset, and develop an improved ceRNA prediction algorithm that learns from known ceRNA interactions and combines currently available prediction methods. Curation of public databases and recent literature yielded 65 experimentally validated ceRNA pairs from three major cancer types (BRCA, PRAD, LIHC). Utilizing public cancer expression datasets from TCGA, we trained a SVM model using the positive examples in the custom database, along with 65 randomly generated negative examples. We named the resulting classifier ceFinder. Five-fold cross-validation of ceFinder yielded a classification accuracy of 77%. To further corroborate our model using an alternative line of experimental evidence, we leveraged a ceRNA database based on High-throughput Sequencing of RNA Isolated by Crosslinking Immunoprecipitation (CLIP-Seq), StarBase. ceFinder separated CLIP-Seq validated positive pairs from StarBase and random negative examples. The receiver operating characteristic (ROC) curve reported an area-under-curve (AUC) score of 0.88. By selecting the optimal model that yields the maximal sum of sensitivity and specificity in the validation data, ceFinder achieved a sensitivity of 0.88 and a specificity of 0.90 in validation. In conclusion, we improved novel ceRNA prediction using supervised learning, and developed a pipeline that can aid the study of disease-relevant regulatory mechanisms in lncRNA biology.