Abstract

Spiro Stilianoudakis

Studies using chromosome conformation capture techniques have shown that the three-dimensional (3D) structure of the genome is tightly compacted into distinct compartments, known as topologically associating domains (TADs), where regulatory elements and genes tend to interact more frequently. This suggests that the boundaries of TADs may play a role in restricting the function of elements such as enhancers, thereby impacting the transcription of genes. The ability to predict which genomic features are most associated with the development of TAD boundaries will allow us to further understand the regulatory role of the 3D structure of the genome. Few methods have been developed to study the role of specific sets of these features on the folding of chromosomes. However, many widely used methods ignore key characteristics of the data that may hinder the performance of certain parametric models. We propose a novel data pre-processing pipeline that addresses irregularities present in genomic data as well as a nonparametric random forest model used for predicting which key molecular drivers are most associated with TAD boundaries. Among the irregularities, we account for heavily imbalanced classes, un-normalized predictors, and large numbers of potentially correlated features. Our model was then compared to established parametric models that did not account for data pre-processing. It was found that the random forest model, when applied to pre-processed data, out performed multiple logistic regression models, with and without LASSO-regularized coefficients. Likewise, the variable importance of the random forest yielded more stable results compared to the coefficients of the multiple logistic regression models.