Predictive modeling using epigenomic annotations

Spiro Stilianoudakis[1] , Mikhail G. Dozmorov[1]

[1] Dept. of Biostatistics, Virginia Commonwealth University, Richmond, VA, 23298, USA

Background. Chromosome conformation capture sequencing technologies have shown that the three-dimensional (3D) structure of the genome is folded into distinct compartments, known as topologically associated domains (TADs) - units of coordinated gene expression. The location of TAD boundaries is highly conserved, suggesting the presence of epigenomic marks aiding in TAD formation. The ability to predict which epigenomic features are most associated with the TAD boundaries will allow to better understand the regulatory role of the 3D structure of the genome.

Existing methods for predicting associations between genomic elements ignore key characteristics of the data. Specifically, the number of TAD boundaries is much less than the number of other genomic regions, leading to heavily imbalanced classes. Furthermore, most methods utilize direct overlap as a means to quantify the association, while distance, the measure of spatial relationships, remains unaccounted for. Consequently, distances on a genomic scale vary widely, leaving uncertainty how the heavily right-tailed distribution of distance measures will affect the model's performance.

Methods. We proposed a novel data pre-processing pipeline that addresses those shortcomings. It includes an elastic-net regularized classification model, where the mixing proportion, α, and the penalization parameter, λ, are tuned for over a grid of values. A number of classifier performance metrics were assessed, including the F1 measure and Matthew Correlation Coefficient (MCC).

Results. Data preprocessing (log2-transformation and standardization) improved the performance of the models. The elastic-net model outperformed multiple logistic regression models, with and without LASSO-regularized coefficients. Likewise, the variable importance of the elastic-net yielded more stable and biologically meaningful results compared to the variables selected by the multiple logistic regression models.

Conclusions. Current methods used to model the epigenomic features associated with TAD boundaries are not robust to handle properties of genomic data. Models applied to unprocessed data can have poor predictive performances. Typical performance assessment metrics, such as AUROC, can mask poor performance of the models; thus, the use of more balanced metrics, such as F1 and MCC, is warranted. The elastic-net model, with its dual use of the l_1 and l_2 regularization terms allows for both the removal of uninformative predictors, while being able to handle multiple correlated epigenomic features. Our model results in better performances and more accurate identification of important features associated with the formation of TAD boundaries.