Enhancer prediction using pattern recognition of epigenetic signals
-> MPRA training

Abstract

Enhancers are an important category of tissue-specific noncoding functional elements, whose activity is often associated with changes in gene expression across different tissues, which are thought to be essential for multi-cellularity. Unfortunately, until recently, enhancers were difficult to characterize experimentally and only a small number of tissue-specific mammalian enhancers were rigorously validated. Hence, for predicting enhancers, it was difficult to train models based on experimentally validated enhancers. Instead, the presence of genomic features associated with enhancers was used to predict them. For example, two of the widest used methods for predicting enhancers were based on the fact that these elements are expected to contain a cluster of transcription factor binding sites and their activity is often correlated with an enrichment of certain post-translational modifications on histone proteins. Now, there are a large number of massively parallel assays for characterizing enhancers. We use the output of these assays to properly train and test a statistical model for predicting enhancers. We initially develop linear filters to identify the occurrence of promoter and enhancer-associated patterns within different epigenetic signals. We then combine these linear filters using simple statistical models these linear features together with few parameters. This statistical model characterizes enhancers in a cell-type specific manner and we show that this model can be transferred without change between various cell lines and even between different organisms. This statistical model allows us to characterize enhancers on a large scale across many tissues and cell lines. It will also allow us to characterize enhancers in cell lines with many experimentally measured transcription factor binding sites and this in turn allow us to see a distinct difference between the type of transcription factor binding at enhancers and promoters, allowing us to construct a secondary model that better discriminates between these two active regulatory regions.