Predicting Tissue Specific Enhancer Activity from Epigenetic Marks and Sequence

Predicting the function of different regions of the genome is one of the grand challenges of genomics. In this project, you will use an existing database of tested enhancers to devise a strategy for predicting novel enhancers, using both sequence features and numerous epigenetic marks. The advent of ChIP-seq has empowered high resolution genome-wide identification of the regions with specific marks, and correlation with functional annotation has identified combinations of marks that are characteristic of active and repressed genes, enhancers, repressors, and other types of functional elements (Ernst et al 2011 "Mapping and analysis of chromatin state dynamics in nine human cell types."). Here, you will be leveraging assays of many different marks in many different tissues to predict enhancers. In doing so, you should also identify which marks, or combinations of marks, are most discriminative in predicting enhancers.

Datasets and Approach

- You will first need a set of reference elements for training, validation, and testing. The Vista Enhancer Browser (http://enhancer.lbl.gov/frnt_page_n.shtml) consists of 1857 sequences from the human and mouse genomes that have been tested in a transgenic mouse enhancer assay. The resulting embryos have been scored for driving expression in the heart, brain (hindbrain, midbrain, forebrain), neural tube, and limb. Use the provided script to parse the database.

- Implement your own version of k-mer enhancer prediction, as in "Discriminative prediction of mammalian enhancers from DNA sequence" (http://www.ncbi.nlm.nih.gov/pubmed/21875935). Predict general enhancer function, tissue-specific enhancer function (heart, limb, brain, etc.), and fine-grain (different regions of the brain) enhancer function. How well does your algorithm predict (as measure by percent correct, precision/recall, AUROC, and any other methods you devise to evaluate your learning algorithm)? Which features are most discriminative in predicting the different enhancers?

- Extend the methods above using epigenetic data from the Roadmap Epigenomics Project (http://www.roadmapepigenomics.org/data). Download the .wig files from the linked GEO accession entries. Using the provided `wigToBigWig` and `bigWigAverageOverBed`, summarize the wiggle tracks over each Vista element. What are the answers to the questions posed above.

Suggested Checkpoint Goals

- Complete the first two bullet points above