genetic sequence

ChIP-seq core experimental resources to understand genome-wide epigenetic interactions and idetify cancer. Preasent irregular noise and bias on various levels. Almoast impossible for human manual to inspect the peaks.

- CNN achieving human-like classifiation accuracy -> supervised learning approach for identifying ChIP-seq peak.
- Data used labeled by human researchers -> annotate the persence or absence of peaks -> trainings data -> predict peaks in previously unseen genomic segments from multiple ChIP-seq datasets.

data description

data in BAM format fromthe ENCODE data portal multible ChIP-seq such as:

- H3K36me3, H3K4me3, H3K27me3, H2AFZ and H3K9ac transcription factor binding such as:
- GATAD2, POLR2A, SMARCE1, and MAX in cancer cells s K562, A549, HepG2, HEK293, and GM12878 addidional input layer for CNN:
- leukemia cell line K562 and (RefSeq) in NCBI RefSeq data include protein-coding locations and pseudogenes.

data preprocessing

model to read BAM files and convert it to vectors with right shape for CNN -> bins 12000 also smoothing and reduce noise -> used max-pooling and Gaussian filter

Results

downloaded 16 ChIP-seq datasets and one ATAC-seq in BAM format -> 3294 genomic segments 66% on avarage are near form the transcription start sites (TSS) of RefSeq genes

Disscussion

Offten used ChIP-seq -> high false pos of peaks