Untitled

This readme file is for how to train PAR/NAR model from Chip-seq peaks and use trained model to scan enhancer sequences for PAR/NAR

0. train TREDNet model using coordinates of enhancers:
   /data/Dcode/common/CenTRED_for_Mehari_94biosamples/TREDNet

   0.1. in submit_local_jobs.sh, train a model on H1 enhancers, with input file located in
   /data/Dcode/common/CenTRED_for_Mehari_94biosamples/input_training_trednet,
   the complete input file is located in
   /data/Dcode/common/CenTRED/hg38/green_celllines/CenTRED_training_files,
   for simplicity, here for HepG2, you can just read dataset in h5 format:
   /data/Dcode/common/CenTRED/hg38/green_celllines/CenTRED_models/BioS11/phase_two_dataset.hdf5

   0.2. the output model will be stored in
   /data/Dcode/common/CenTRED_for_Mehari_94biosamples/CenTRED_models/part1
1. train PAR/NAR model: /data/Dcode/common/CenTRED_for_Mehari_94biosamples/PARNNAR_model
   /data/Dcode/gaetano/CenTRED/CenTRED_for_PARNARs/PARNNAR_model
   1.1. FIMO predicted motif positions as "true TF binding sites" in the directory:
   /data/Dcode/common/CenTRED_for_Mehari_94biosamples/PARNNAR_model/FIMO_identified_Chipseq_TFBS
   file total_final_chip2fimo_HepG2.pvaluee_04.merged is from FIMO scanned motif location in
   all HepG2 TF Chip-seq peaks and combined together,
   eg, HNF4A motif position located in HNF4A Chip-seq peaks.
   These motif locations will be the positive sets for PAR/NAR model training

   ```
   1.2. generate input positive/control sets for PAR/NAR model training:
   /data/Dcode/common/CenTRED_for_Mehari_94biosamples/PARNNAR_model/step1_input_PARNNAR
   in submit_step0_inputfile.sh, overlaping HepG2 enhancer with motif locations in previous
   step to get the positive sets in HepG2 enhancer.
   Control sets are generated in HepG2 enhancer regions excluding the motif locations,
   the output file will be: list_control_in_enhancer.bed and list_motif_in_enhancer.bed

   in submit_step1_genebasepair_bychrom.sh,  generating 220 features for each of the nucleotides
   in positive/control sets, individually for peak/dip (require precalculated normalized
   deltascore for each of the nucleotide, see below section 2.1, the input format is here:
   /data/Dcode/common/CenTRED_for_Mehari_94biosamples/gene_mutagenesis/output_deltascore/BioS11_
   1kb/output.txt.total.BioS11.fpr5.normscore.newformat). The output files:
   list_control_in_enhancer.bed.withenh.feature.chr1,
   list_motif_in_enhancer.bed.withenh.dip.feature.chr1,
   list_motif_in_enhancer.bed.withenh.peak.feature.chr1

   in submit_step2_split.sh, combined all files with 220 features and split them into training
   (none chrom 8 and 9) and testing sets (chrom 8 and 9). The output files are:
   input_peak.train, input_peak.test, input_dip.train and input_dip.test

   1.3. train the PAR/NAR model using the training and testing sets from previous steps:
   /data/Dcode/common/CenTRED_for_Mehari_94biosamples/PARNNAR_model/step2_train_PARNNAR. The
   output files: BioS11_HepG2hg38_peak
   ```

2. using already trained PAR/NAR model to scan DNA sequences (eg. HepG2 or other tissue enhancers)
   2.1. generate in-silico mutagenesis for enhancers in
   /data/Dcode/common/CenTRED_for_Mehari_94biosamples/gene_mutagenesis/step1_gene_mutagenesis/ (note

that this step is also required for the PAR/NAR model training in previous section 1.2, for the purpose of PAR/NAR model training, we can consider enhancers with fpr>0.05 only.)
in submit_step1_fasta_allenh.sh get the fasta sequence for the enhancers
in submit_step2_run_trednet.sh use already trained TREDNet enhancer model (here HepG2 model, trained in previous section 0.2,
/data/Dcode/common/CenTRED_for_Mehari_94biosamples/CenTRED_models/part1/BioS11.phase_two.hg38) to generate raw deltascore
in submit_step3_calculate_deltascore.sh generate normalized deltascore

2.2. again generate 220 features for each nucleotide in the enhancers or the DNA sequence you want, in /data/Dcode/common/CenTRED_for_Mehari_94biosamples/gene_mutagenesis/step2_gene_220feature/submit_step1_220feature.sh, this scrip will copy the ./original_code to generate a separate directory for each tissue and generate 220 features only for enhancers in that tissue, example is:
/data/Dcode/common/CenTRED_for_Mehari_94biosamples/gene_mutagenesis/step2_gene_220feature/feature220_BioS94_1kbp/list_allenh_in_enhancer.bed.withenh.feature.1, please note that this step will generate a lot of data with large size, don't forget to delete them after you prediced PAR/NAR.

2.3. generate PAR/NAR by scanning the 220 features from section 2.2, in
/data/Dcode/common/CenTRED_for_Mehari_94biosamples/gene_mutagenesis/step3_gene_peakNdip
in submit_step1_scan_peakNdip.sh, using model from
/data/Dcode/common/CenTRED_for_Mehari_94biosamples/PARNNAR_model/step2_train_PARNNAR/BioS11_HepG2hg38_peak (previous section in 1.3) to predict peak/dip status for each nucleotide, the output file is located in:
/data/Dcode/common/CenTRED_for_Mehari_94biosamples/gene_mutagenesis/output_peakNdip/peak/feature220_BioS11_1kbp/output.BioS11.1
in submit_step2_filter_fpr.sh, filter these peak/dip prediction by fpr 0.01 or fpr 0.05 based on PAR/NAR model, the output file is
/data/Dcode/common/CenTRED_for_Mehari_94biosamples/gene_mutagenesis/output_peakNdip/peak/feature220_BioS11_1kbp/output.BioS11.total.peak.fpr1.dis_ease.sorted
in submit_step3_gene_PASNDAS.sh, merge the filtered peak/dip nucleotides to form PAR/NAR regions.

3. instead of using top/bottom 5% deltascore to define PAR/NAR directly, you can also build an extra layer of PAR/NAR model based on top/bottom 5% deltascore is in
/data/Dcode/common/CenTRED/hg38_PASNDAS/step5_DLPARNAR_ontop5percent