TREDNet PARNARs

This document outlines the process of training the PAR/NAR model using ChIP-seq peaks and utilizing the trained model to scan enhancer sequences.

Step 0: Train the TREDNet Model with Enhancer Coordinates

0.1 Navigate to the Training Directory

Go to the path:

```
/data/Dcode/gaetano/CenTRED/CenTRED_for_PARNARs/TREDNet
```

0.2 Submit the Training Job

Run the following script to start the training:

```
sh submit_local_jobs.sh H1
```

This trains the model using data from **H1 enhancers**. Replace `H1` with any available biosample from the input files directory.

0.3 Input Files

The input data is located at:

```
/data/Dcode/gaetano/CenTRED/CenTRED_for_PARNARs/input_training_trednet
```

The complete input file can be found at:

```
/data/Dcode/common/CenTRED/hg38/green_celllines/CenTRED_training_files
```

For simplicity, a pre-processed dataset for **HepG2** is available in HDF5 format:

```
/data/Dcode/common/CenTRED/hg38/green_celllines/CenTRED_models/BioS11/phase_two_dataset.hdf5
```

0.4 Output

The trained model will be saved in:

```
/data/Dcode/gaetano/CenTRED/CenTRED_for_PARNARs/CenTRED_models/part1
```

Step 1: Training the PAR/NAR Model

1.1 Define Positive TF Binding Sites

FIMO-predicted motif positions serve as **true TF binding sites** and are stored in:

```
/data/Dcode/gaetano/CenTRED/CenTRED_for_PARNARs/PARNNAR_model/FIMO_identified_Chipseq_TFBS
```

- The file `total_final_chip2fimo_HepG2.pvaluee_04.merged` contains motif locations scanned by FIMO across all **HepG2 TF ChIP-seq peaks**.
- For example, **HNF4A motif positions** are identified within **HNF4A ChIP-seq peaks**.
- These motif locations will be used as **positive training sets** for the PAR/NAR model.
  ToDo: Make sure you have the file `.merged` in the given folder. Therefore, add a new file for a different cell-line.

1.2 Generate Input Positive and Control Sets

Go to the directory:

```
/data/Dcode/gaetano/CenTRED/CenTRED_for_PARNARs/PARNNAR_model/step1_input_PARNNAR
```

1.2.1: Create Positive and Control Sets

Run:

```
sh submit_step0_inputfile.sh
```

This script will:

- Overlap **HepG2 enhancers** with **motif locations** from Step 1.1 to create **positive sets** (FIMO across all **HepG2 TF ChIP-seq peaks**).
- Generate **control sets** from **HepG2 enhancer regions** that exclude motif locations.

Output Files:

- `list_control_in_enhancer.bed` (control sets)
- `list_motif_in_enhancer.bed` (positive sets)
  *in the same folder.

1.2.2: Extract Features for Each Nucleotide

Run:

```
sh submit_step1_genebasepair_bychrom.sh
```

This generates **220 features** per nucleotide in the **positive and control sets** (separately for peak/dip regions).

Required Input: Precomputed **normalized delta scores** per nucleotide.

```
/data/Dcode/common/CenTRED_for_Mehari_94biosamples/gene_mutagenesis/output_deltascore/BioS11_1kb/output.txt.total.BioS11.fpr5.normscore.newformat
```

Output Files (for chromosome 1 as an example):

- `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`

1.2.3: Split Data into Training and Testing Sets

Run:

```
sh submit_step2_split.sh
```

This script will:

- **Merge feature files** containing 220 features.
- **Split data** into training and testing sets:
  - **Training set:** Excludes chromosomes **8 and 9**
  - **Testing set:** Includes chromosomes **8 and 9**

Output Files:

- `input_peak.train`, `input_peak.test`
- `input_dip.train`, `input_dip.test`

1.3 Train the PAR/NAR Model

Directory:

```
/data/Dcode/common/CenTRED_for_Mehari_94biosamples/PARNNAR_model/step2_train_PARNNAR
```

Using the training and testing sets from Step 1.2, train the **PAR/NAR model**.

- `input_peak.train`, `input_peak.test`
- `input_dip.train`, `input_dip.test`

Output File:

```
BioS11_HepG2hg38_peak
```

Step 2: Using a Pre-trained PAR/NAR Model to Scan DNA Sequences (e.g., HepG2 or Other Tissue Enhancers)

2.1. Generate In-Silico Mutagenesis for Enhancers

- **Directory:**
  = `/data/Dcode/gaetano/CenTRED/CenTRED_for_PARNARs/gene_mutagenesis/step1_gene_mutagenesis/`
- This step is also required for **PAR/NAR model training** (see Section 1.2).
- For **PAR/NAR model training**, only consider **enhancers with FPR > 0.05**.

Steps to Process Enhancers Using the Pre-trained Model

1. **Extract Fasta Sequences for Enhancers**
   - **Script:** `submit_step1_fasta_allenh.sh`
   - Retrieves the **fasta sequences** for enhancers.
   - Output in folder: `${EID}_${length}`
2. **Generate Raw Delta Score Using TREDNet**
   - **Script:** `submit_step2_run_trednet.sh`
   - Uses a **pre-trained TREDNet enhancer model** (e.g., HepG2, trained in Section 0.2).
   - **Model Path:**
     `/data/Dcode/gaetano/CenTRED/CenTRED_for_PARNARs/CenTRED_models/part1/BioS11.phase_two.hg38`.
3. **Normalize Delta Scores**
   - **Script:** `submit_step3_calculate_deltascore.sh`

- Computes the **normalized delta scores**.

2.2. again generate 220 features for each nucleotide in the enhancers or the DNA sequence you want

The script `submit_step1_220feature.sh` (located at:
📂 `/data/Dcode/gaetano/CenTRED/CenTRED_for_PARNARs/gene_mutagenesis/step2_gene_220feature/`)
automates the generation of 220 features for enhancers in different tissues.

How It Works:

- It **copies** the contents of `./original_code` to create a **separate directory** for each tissue.
- It then **generates 220 features only** for enhancers in that tissue.

Example Output Directory:

📂
`/data/Dcode/common/CenTRED_for_Mehari_94biosamples/gene_mutagenesis/step2_gene_220feature/feature220_BioS94_1kbp/`

📄 Example feature file:
`list_allenh_in_enhancer.bed.withenh.feature.1`

⚠ Important Note:

- This step generates **a large amount of data**.
- **Make sure to delete** unnecessary files after predicting PAR/NAR to free up storage.