TREDNet PARNARs

Training and Utilizing the PAR/NAR Model with ChIP-seq Peaks

Overview

This document outlines the process of training the PAR/NAR model using ChIP-seq peaks and applying 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
```

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

0.3 Input Files

Input data location:

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

Complete input file:

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

Pre-processed dataset for HepG2 (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**:

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

- Example: `total_final_chip2fimo_HepG2.pvaluee_04.merged`
- **ToDo:** Ensure `.merged` file exists for each cell line.

1.2 Generate Input Positive and Control Sets

Navigate to:

```
/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
```

- Generates **positive sets** (FIMO motif locations in HepG2 enhancers).
- Generates **control sets** (HepG2 enhancers excluding motif locations).

1.2.2 Extract Features per Nucleotide

Run:

```
sh submit_step1_genebasepair_bychrom.sh
```

- Computes **220 features** per nucleotide.

1.2.3 Split Data into Training and Testing Sets

Run:

```
sh submit_step2_split.sh
```

- **Training set:** Excludes chromosomes **8 and 9**.
- **Testing set:** Includes chromosomes **8 and 9**.

1.3 Train the PAR/NAR Model

Navigate to:

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

Output File:

```
BioS11_HepG2hg38_peak
```

Step 2: Scanning DNA Sequences Using a Pre-trained PAR/NAR Model

2.1 Generate In-Silico Mutagenesis for Enhancers

Navigate to:

```
/data/Dcode/gaetano/CenTRED/CenTRED_for_PARNARs/gene_mutagenesis/step1_gene_mutagenesis/
```

- **Extract fasta sequences** using:

```
sh submit_step1_fasta_allenh.sh
```

- **Generate raw delta scores** using TREDNet:

```
sh submit_step2_run_trednet.sh
```

- **Normalize delta scores:**

```
sh submit_step3_calculate_deltascore.sh
```

2.2 Generate 220 Features for Each Nucleotide

Navigate to:

```
/data/Dcode/gaetano/CenTRED/CenTRED_for_PARNARs/gene_mutagenesis/step2_gene_220feature/
```

Run:

```
sh submit_step1_220feature.sh
```

2.3 Scan DNA Sequences Using the Pre-trained Model

Navigate to:

```
/data/Dcode/gaetano/CenTRED/CenTRED_for_PARNARs/gene_mutagenesis/step3_gene_peakNdip
```

Run:

- **Predict peak/dip status:**

```
sh submit_step1_scan_peakNdip.sh
```

- **Filter predictions by FPR (0.01 or 0.05):**

```
sh submit_step2_filter_fpr.sh
```

- **Merge filtered peak/dip nucleotides into PAR/NAR regions:**

```
sh submit_step3_gene_PASNDAS.sh
```

Step 3: Alternative Approach - Using Delta Scores for PAR/NAR Modeling

Instead of defining PAR/NAR directly using top/bottom 5% delta scores, an additional modeling step can be performed. Navigate to:

```
/data/Dcode/common/CenTRED/hg38_PASNDAS/step5_DLPARNAR_ontop5percent
```

This approach builds an extra layer of the PAR/NAR model based on delta scores.

Summary

- **Step 0:** Train TREDNet with enhancer coordinates.
- **Step 1:** Train PAR/NAR using motif locations from ChIP-seq peaks.
- **Step 2:** Scan sequences using the trained model and generate predictions.
- **Step 3:** Alternative approach using delta scores.

Ensure **proper storage management** to handle large data files efficiently.