Data Preparation Guide- CH3 project

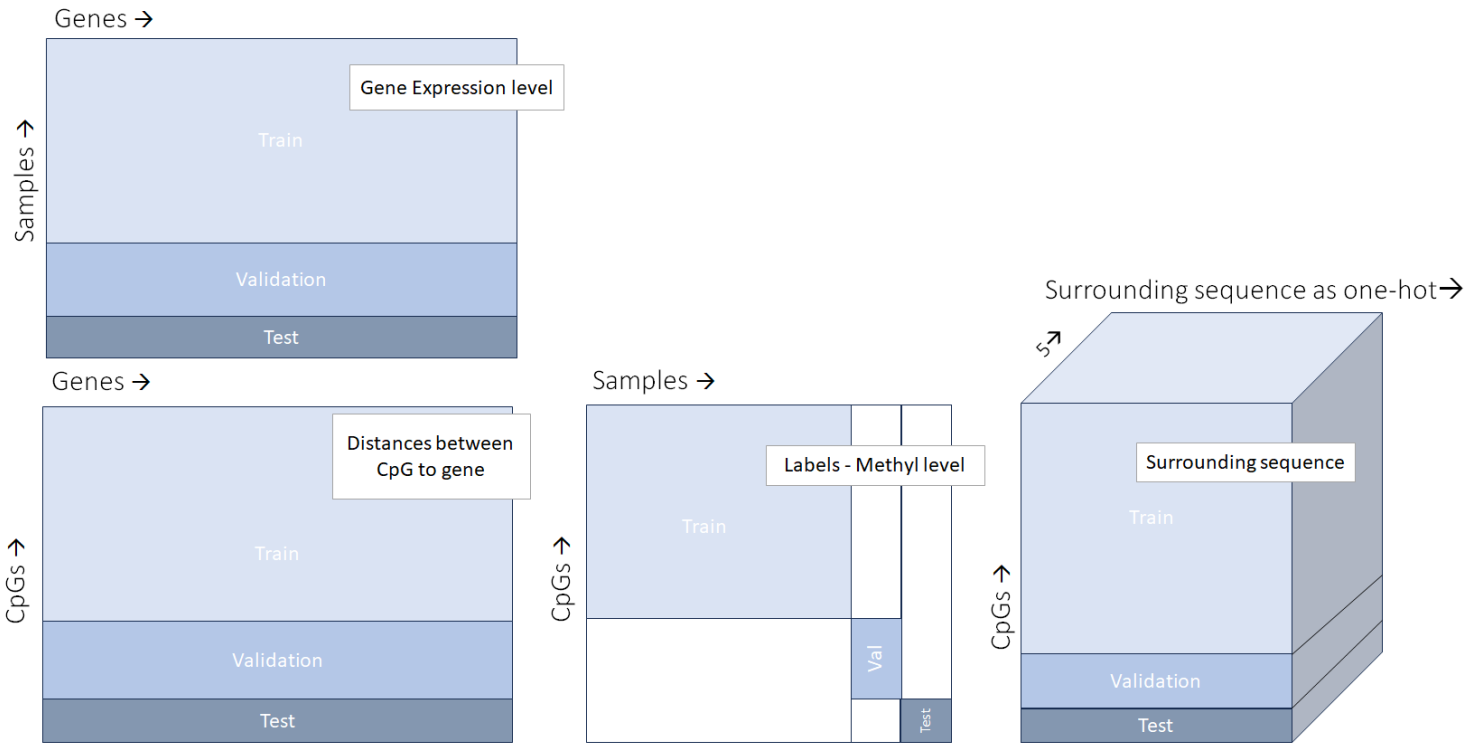
Adi Gotliber, Dana Keydar

Overview:

Inputs:

- **Specific per subject data:** BRCA, LUAD methyl levels and gene expression per sample
- **General data:**
  - Human genome data
  - CpG locations probes
  - Gene location data

Output:



- **Gene expression** per gene per subject (#Genes X #samples) - not normalized
- **Labels**- methyl level per cpg per subject (#CpG X #samples)
- **One hot for surrounding sequence** per CpG (#CpG X 4000) (4000 = 5 base x 800 surrounding bases)
- **Distances**- distance of each CpG from each gene (#CpG X #Genes). CpGs sites within x base-pairs from the nearest gene
  - Optional x values : 2000 (model 1) , 10,000 (model 2), No limit (model 3)

- Distances are being calculated only for genes and CpGs from the same chromosoms.

Data prep notebooks:

- One hot preparation: get_seq_for_cpgs.ipynb
- Labels, expressions and distances file preparation : get_lables_expressions_and_distances.ipynb
- Data sampling - while using model 1- we need to align the CpGs in the labels file and in the seq file, to the ones in the distance file.

Data prep process and code:

You can have a look at the training files here : sampled_files_creation.ipynb

Input files for the entire data prep:

1. **Human Genome data** -
   - Source: open source data.
     Link to download (downloading the data for each chromosome)
   - Directory in our project google drive.
     dir_hg19 =
     `"/content/drive/MyDrive/MSC/CH3_Project/res/hg_chromosoms"`
     Saved three chromosomes as example, to re-prepare, need to re-download all of the chromosomes from the Link, fa.gz files.
     Description: This file contains the chromosome sequence.

2. **CpG locations file**-
   - Source- Xavier's data files
   - Directory in our project google drive. dfMethyl_path =
     `"/content/drive/MyDrive/MSC/CH3_Project/res/450k_probes_ChroMM.bed"` (This path appears in the ipynb).
   - Directory in Yachini's lab google cloud
   - Description: This file contains the start and end location of each CpG. We used it to create the sequence around cpg (for the one_hot_surrounding_sequence file)

3. **Raw data of BRCA and LUAD samples**
   - Source- Xavier's data files
   - Directory- in out project's g.drive
     i. raw_data_brca_path =
     `'/content/drive/MyDrive/MSC/CH3_Project/Spatial Methyl/raw_data/BRCA.RData'`
     ii. raw_data_luad_path =
     `'/content/drive/MyDrive/MSC/CH3_Project/Spatial Methyl/raw_data/LUAD.RData'`

- Description: The raw data contains expression data and methyl data. Expression df contains the expression per gene per sample. The Methyl df contains the methyl level for each cpg.
- Preparation: Merging BRCA and LUAD df's.

4. **Genes locations** - genomic.gbff
   - Source- Open source
     Link for download-
     https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000001405.37/
   - Path in our project -
     `"/content/drive/MyDrive/MSC/CH3_Project/res/genomic.gbff"`
   - Description: The genes locations (start-end). Used to get the distances between each CpG to genes.

Prepare the training files : (by stages)

One hot surroundingSeq preparation: get_seq_for_cpgs.ipynb

1. **Probe/CpG to surroundingSeq file**
   a. `"/content/drive/MyDrive/MSC/CH3_Project/res/hg_chromosoms/probeToSurroundingSeqFilePrefixAll/probe_to_surroundingSeq_.csv"`
   b. Description : A table that maps the probe to its surrounding sequence in the DNA.
   c. Preparation Extracted the sequence surrounding the cpg from the human genome file.
   d. Relevant functions: `getSurroundingSeqTablePerChr`
2. **SurroundingSeq to one-hot-encoding file**
   a. Description: Encoding the probe surrounding sequence of each cpg into a one hot representation.
   b. Relevant function : `sequences_to_one_hot_mtrx`
   c. Drive path :
   `"/content/drive/MyDrive/MSC/CH3_Project/res/probe_to_surrounding_seq_one_hot_formatted.csv"` **to change to google cloud**

Distances file preparation : get_lables_expressions_and_distances.ipynb

1. **Gene_to_pos map:** Mapping the genes in the raw expression data to their positions of their coding sequences in the genome (start and end positions per gene). For this mapping we used `"genomic.gbff"` that includes the positions of the coding sequences in the genome.

   a. Path - `"/content/drive/MyDrive/MSC/CH3_Project/for_distances/geneToPos.csv"`
   b. Function- `"createGenePositionsDict"`
2. **Probe_to_pos map:** Mapping the Probe/CpG of interest to their positions, using the `"450k_probes_ChroMM.bed"` file that contains the start and end location of each CpG.
   To change to **CpG_to_pos**
   a. Path- `("/content/drive/MyDrive/MSC/CH3_Project/for_distances/probe_to_pos.csv")`
   b. Function- `"createProbePositionsDict_adjusted"`
3. Distances -
   a. Path- `"/content/drive/MyDrive/MSC/CH3_Project/res/distances_try2.cs"`
   b. Function- `"createDistanceMatrx_adjusted"`
   c. How does it work? For each CpG, for each gene, if they are from the same chromosome, we save their inverse distance in 1/bp (one over base-pairs). When in different chromosomes or the inverse distance is lower than threshold, it will be set to zero.