Data Preparation Guide- CH3 project

Adi Gotliber, Dana Keydar

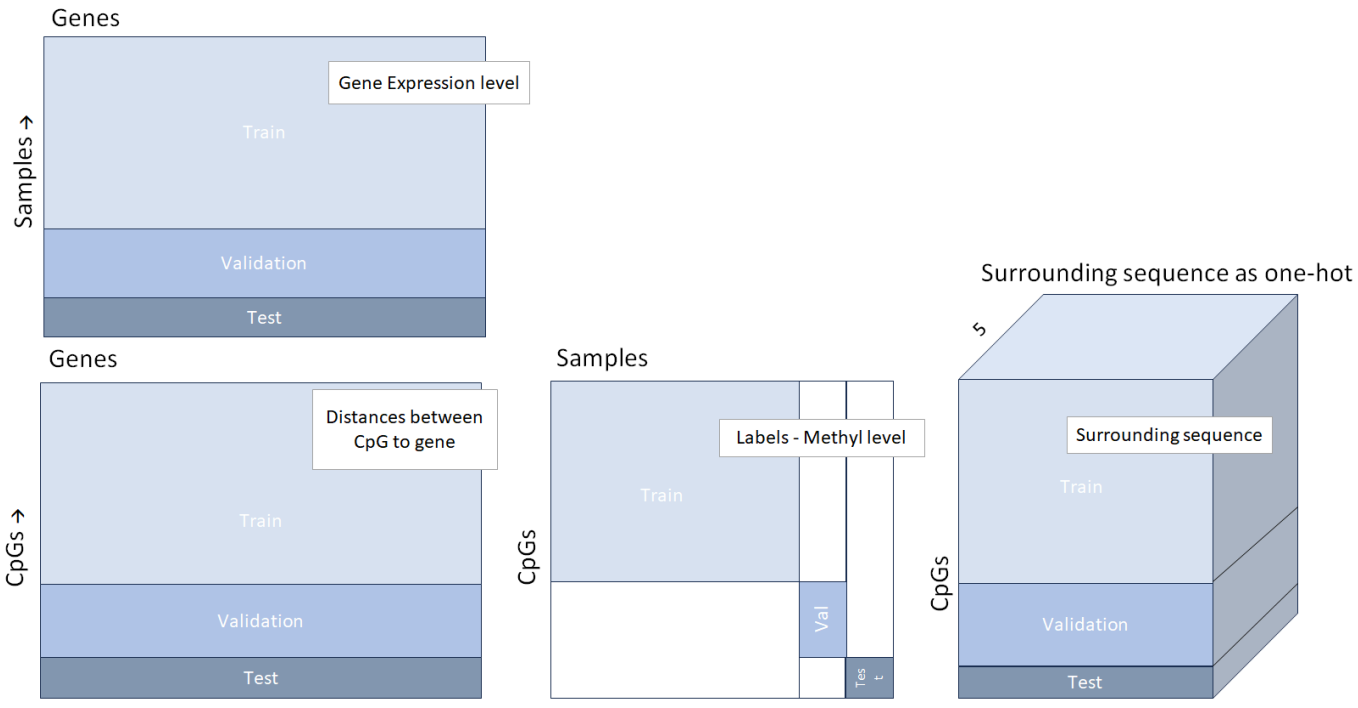
Overview:

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 chromosome.



2. **CpG locations file**-
    - Source- Xavier's data files
    - The data: 450k_probes_ChroMM.bed
    - 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 = /raw_data/BRCA.RData
        ii. Raw_data_luad_path = /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 downloading-
      https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000001405.37/
    - The file - `genomic.gbff`
    - Description: The genes locations (start-end). Used to get the distances between each CpG to genes.

Prepare the training files : (by stages)

Input raw data: google cloud: adi-gotliber-methylation/Raw_data_files
Output data: google cloud: adi-gotliber-methylation/Full_data_for_2k_distance_model_1/

One hot surroundingSeq preparation: Code

1. **Probe/CpG to surroundingSeq file**
    a. File name: 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. File name: probe_to_surroundingSeq_one_hot_formatted.csv.
    b. Description: Encoding the probe surrounding sequence of each cpg into a one hot representation.
    c. Relevant function : sequences_to_one_hot_mtrx

Distances file preparation : code

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. File name: 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.
    a. File name: Prob_to_pos.csv
    b. Function- "createProbePositionsDict_adjusted"
3. **Distances -**
    a. File name: distances.csv
    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.