Model Script Documentation

The "Model" R script is designed to train predictive models using Lasso regression for each gene based on methylation data. It leverages the gene-to-CpG site mapping created in the "Mapping" script to identify relevant CpG sites for each gene and then trains a model to predict gene expression levels from methylation patterns.

Libraries Used

- **glmnet**: For performing Lasso regression with cross-validation.

Input Files

- Gene-to-CpG site mapping file (**gene_to_cpg_site_mapping.rda**): A mapping between genes and their associated CpG sites.
- Expression data file (**expression.csv**): Contains gene expression levels, with genes as rows and samples as columns.
- Methylation data file (**methylation.csv**): Contains methylation levels at specific CpG sites, with CpG sites as rows and samples as columns.

Output File

- Regression model file (**regression.output.rda**): A serialized R data frame saved as an **.rda** file, which contains the Lasso regression models trained for each gene, along with their predictions and R-squared values.

Main Functions

- **train_predict_lasso(gene, cg_sites, expression_data, methylation_data)**: Trains a Lasso regression model for a single gene using its associated CpG sites, predicts gene expression levels, and calculates the R-squared value of the predictions.
- **apply_lasso_to_genes(expression_data, methylation_data, gene_to_cpg_site_mapping)**: Applies the **train_predict_lasso** function to each gene based on the gene-to-CpG site mapping and compiles the results into a data frame.

Usage

1. Set the working directory to the location of the script.
2. Ensure that the required input files (**gene_to_cpg_site_mapping.rda**, **expression.csv**, and **methylation.csv**) are present in the working directory.
3. Run the script. It will load the **glmnet** library, read the input files, and train Lasso regression models for each gene based on the provided gene-to-CpG site mapping.
4. The script saves the trained models, their predictions, and R-squared values into a data frame, which is then serialized and saved as **regression.output.rda**.

Note

This script assumes that the expression and methylation datasets contain the same samples in the same order. Users should verify this alignment before running the script. The gene expression levels and methylation patterns must be properly matched for the models to be trained accurately.