SALMON package (MATLAB)

Heeju Noh (heeju.noh@chem.ethz.ch) and Rudiyanto Gunawan (rudi.gunawan@chem.ethz.ch)

July 19, 2017

SALMON version 1.0

The MATLAB subroutines in the SALMON package (**v.1.0**) have been successfully tested on **MATLAB® 2014b to 2017a** platforms. Please refer to the SALMON manuscript for more detailed information about the algorithm. Any questions regarding SALMON usage can be addressed to heeju.noh@chem.ethz.ch or to rudi.gunawan@chem.ethz.ch.

Installation instruction:

1. Unzip the package ***SALMON_1.0_MAT.zip*** to a preferred folder.
2. Download the MATLAB version of GLMNET package, and unzip the GLMNET package under a new subfolder in SALMON.

3. Set the current working directory to SALMON in MATLAB.
4. Add the path for GLMNET package.

The SALMON package includes the following:

1. example_data

A subfolder in SALMON package, containing microarray data from the chromatin targeting study using mouse pancreatic beta cells [**1**]:

- ***lfc_mouse-pancreas-beta_13010genesX87samples.txt***: log2FC data (*lfc*), pre-processed as described in SALMON manuscript
- ***list_of_genes.txt***:The list of gene symbols corresponding to the rows in the log2FC data
- ***table_of_samples.txt***: The table of sample descriptions including time points (if in time-series) and group indices (same index for the same drug)
- ***edges-TFTG_mouse_pancreas_fromCellNet.txt***: Transcription factor (TF)-gene network for mouse pancreas cells obtained from CellNet database [**2**]
- ***edges_ppi_fromSTRING_short.txt***: Protein-protein interactions for mouse cells obtained from STRING database [**3**]
- ***A.mat***: The matrix containing estimated a_{ij} from the log2FC data, where a_{ij} indicates the estimated regulatory impact on gene i by protein j. This is one of the files produced by running the example script with the mouse data.
- ***Pscore.mat***: The protein scores for each drug in the dataset. This is one of the files produced by running the example script with the mouse data.

2. findiff.m

This function implements a 2nd order accurate finite difference for calculating slopes (time-derivatives) of the log2FC data using three time points. The function is used in *generateSlope* function below.

3. generatePGN.m

This function constructs the protein-gene network (PGN) by combining TF-gene and protein-protein interaction networks.

```
pgn = generatePGN(GList, tftg, ppi, tftg_thre, ptf_thre, ppi_thre)
```

INPUT ARGUMENTS:

- `GList`: The vector of genes in the same order of the genes in log2FC data. The length of `GList` should be the same as the number of rows in log2FC data.
- `tftg`: The matrix of TF-gene interactions. The first column is the list of TFs, and the second column is the list of genes regulated by the corresponding TFs. The third column is optional, and if present, the column should contain the (confidence) score for each interaction.
- `ppi`: The matrix of protein-protein interactions. Each row of the first two columns give the protein pairs with interactions. The third column is optional, and if present, the column should contain the (confidence) score for each interaction.
- `tftg_thre`: A threshold for TF-gene interactions. This variable is used only when the confidence score of TF-gene interactions are given in the matrix `tftg`. Any TF-gene interactions with confidence scores lower than the threshold will be excluded.
- `ptf_thre`: A threshold for protein-TF interaction. This variable is used only when the confidence score of protein-TF interactions are given in the matrix `ppi`. Any protein-TF interactions with the scores lower than the threshold will be excluded.
- `ppi_thre`: A threshold for protein-protein interaction. This variable is used only when the confidence score of protein-protein interactions are given in the matrix `ppi`. Any protein-protein interactions with the scores lower than the threshold will be excluded.

OUTPUT ARGUMENTS:

`pgn`: The adjacency matrix of PGN.

4. generateSlope.m

The function for calculating slope matrix from log2FC data. If more than two time points are available for a given drug/compound treatment, then a 2nd order accurate finite difference approximation is used for calculating the slopes. If only two time points are available, then a linear slope between the two time points is used.

```
slope = generateSlope( lfc, tp, group )
```

REQUIRES: **findiff.m**

INPUT ARGUMENTS: * `lfc`: The matrix of log2FC data. Each row represents a gene and each column represents a sample. * `tp`: A vector of time points of the samples in the matrix `lfc`. The length of the vector should be the same as the number of samples (i.e. the number of columns in the matrix `lfc`). * `group`: A vector of indices indicating the set of samples from a particular drug/compound treatment. The (time-series) samples from the same drug treatment experiment should have the same unique index. The length of the vector should be the same as the number of samples.

OUTPUT ARGUMENTS:

`slope`: the slope matrix having the same dimension as the log2Fc matrix.

5. run_salmon_example.m

An example script of running SALMON for mouse pancreas data.

6. salmon.m

The main function for SALMON for generating the protein scores for each drug treatment.

```
[Pscore, A] = salmon( lfc, slope, pgn, grplist, kfold, par, numCores )
```

INPUT ARGUMENTS:

- `lfc`: The matrix of log2FC data. Each row represents a gene and each column represents a sample.
- `slope`: The slope matrix from log2FC data. This matrix can be obtained using the function *generateSlope()*. If the data are not time-series, set slope to an empty matrix (i.e. `slope`=[]).

- `pgn`: The adjacency matrix of the protein-gene regulation network. This matrix can be created using the function *generatePGN()*.
- `grplist`: The group index for protein scoring. This vector defines the samples for which the protein scores are computed. The length of this vector should be the same as the number of samples in the log2FC matrix. A single (aggregate) protein score is generated for samples with the same index. The group indices should be a consecutive integer starting from 1 to the number of groups.
- `kfold`: The number of folds used in the k-fold cross validation.
- `par`: A Boolean variable *TRUE* or *FALSE* indicating whether to use parallel computing. The default is *FALSE* (no parallel computation).
- `numCores`: The number of CPU cores to be used for parallel computing. This parameter is considered only if par is *TRUE*. The default is 4. OUTPUT ARGUMENTS:
- `Pscore`: The matrix of protein scores. Each row corresponds to a gene following the same order as the one in the log2FC data, while each column corresponds to a group of samples as defined in the `grplist`.
- `A`: The matrix containing estimated a_{ij} from the log2FC data, where a_{ij} indicates the estimated regulatory impact on gene i by protein j. The rows correspond to genes having at least one regulator based on the PGN (i.e. zeros for the others).

REFERENCES:

[1] Kubicek, S., J. C. Gilbert, D. Fomina-yadlin, A. D. Gitlin, and Y. Yuan. 2012. Chromatin-targeting small molecules cause class-specific transcriptional changes in pancreatic endocrine cells.

[2] Cahan, P., H. Li, S. A. Morris, E. Lummertz Da Rocha, G. Q. Daley, and J. J. Collins. 2014. CellNet: Network biology applied to stem cell engineering. Cell 158 (4): 903-915.

[3] Szklarczyk, D., A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, et al. 2015. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Research 43 (D1): D447-D452.