Deblender v1.0

1. INSTALLATION

The current version of Deblender requires Matlab version > 2012b. "Statistics and Machine Learning", "Optimization" and "Parallel Computing" toolboxes should be installed. Copy and paste the selected folder Deblender into an active MATLAB path.

2. FORMAT OF DATA

1. Load the expression data into the workspace (class double) as positive raw (non-logged) values after normalization, without zeros (add small offset, if needed) or NaNs. If data is logged, transform as "2.^data".
2. Do not filter genes before applying Deblender but ideally use one probe per unique gene identifier.
3. The number of mixed samples should be at least 2 (important for underdetermined cases).
4. In case the mixed dataset has replicates you can either average before analysis or proceed with replicates.
5. Load separately the gene names (or probes) of mixed expression data or marker gene lists (or probe lists) as textdata or numeric. Ensure that all names are unique and the marker names match the nomenclature system of genes in the mixed dataset.
6. In case of using marker gene lists, ensure that there is at least one marker gene for each cell type you will examine in the mixture.

   Notation: X: mixed data, S: cell-specific expression data, A: proportions

3. TYPES OF DATA ANALYSIS

Deblender has 5 pipelines:

"semi-supervised_overdetermined": it is meant for mixed data where the user has known marker gene lists for all cell types (semi-supervised) involved in the mixture along with the exact number of participating cell types. Call "calc_A_known_marker_genes.m" and check parameters if you want to calculate the mixture proportions based on DSA (stage I) or on NMF-CELLMIX using as initialization the DSA result in one of the replicated experiments (Stage I and Stage II). Then use the "calculate_S.m" to estimate the cell-specific gene expression data and 'calc_standard_error.m' to calculate standard error for each gene and metrics for goodness of fit. For details about the parameters of each function, check the comments in the respective functions.

"semi-supervised_underdetermined": it is meant for mixed data where the user has known marker gene lists for all cell types (semi-supervised) involved in the mixture along with the exact number of participating cell types but the number of mixed samples is lower than the number of cell types. For underdetermined cases, only mixture proportions can be calculated. Call "calc_A_known_marker_underdetermined.m" and check the parameters to select one of three adapted

NMF schemes. For details about the parameters of each function, check the comments in the respective functions.

"Unsupervised_overdetermined": it is meant for mixed data where the user does not have any prior information about marker genes about the mixture dataset but knows the exact number of participating cell types. Call "calc_A_unsupervised.m" and check the parameters to select the option to calculate the mixture proportions based only on DSA (stage I) or if you want to calculate the mixture proportions based on NMF with the DSA result as initialization in one of the replicated experiments (stage I and stage II). Then, call "calculate_S.m" to estimate the cell-specific gene expression data and 'calc_standard_error.m' to calculate standard error for each gene and metrics for goodness of fit. For details about the parameters of each function, check the comments in the respective functions.

"Unsupervised_underdetermined": it is meant for mixed data where the user does not have any prior information about marker genes about the mixture dataset but knows the exact number of participating cell types and the number of mixed samples is lower than the number of participating cell types. For underdetermined cases, only mixture proportions can be calculated. Call "calc_A_unsupervised_underdetermined.m" and check the parameters to select one of three adapted NMF schemes. For details about the parameters of each function, check the comments in the respective function.

"MDL_calc": it is meant for mixed data for which the proportion estimation refers to overdetermined cases, where the user does not have any prior information about marker genes about the mixture dataset and does not know the exact number of participating cell types. In this case, call "calc_MDL.m" to examine Minimum Description Length (MDL) criterion in a range of involved cell types {2, N} and check where the MDL reaches its minimal value. Of note, by default (stage I&II) is used for calculating proportions (if you want to change check the comments in the function). After appointing the right number of involved cell types, call "calc_A_unsupervised.m" to calculate mixture proportions and call "calculate_S.m" to estimate cell-specific gene expression profiles and 'calc_standard_error.m' to calculate standard error for each gene and metrics for goodness of fit.

4. NUMERICAL SOLVERS

Deblender employs the following solvers:

1. Lsqnonneg (Matlab function)
2. Lsqlin (Matlab function)
3. Quadprog (Matlab function)
4. Non-Negative matrix factorization (NMF) ('nnmf' Matlab function) and adapted NMF schemes (check manuscript). Of note, we provide the user the option to normalize the final estimated W and H factors as recommended by the 'nnmf' Matlab function (i.e. normalize first H and then W so that the objective function value remains unchanged) or by normalizing first W and then H or no normalization. To change this option, check the comments of the functions.
5. Unified Particle Swarm Optimization (UPSO): The UPSO has been implemented based on "Parsopoulos KE , Vrahatis MN. Parameter selection and adaptation in Unified Particle Swarm

Optimization. Mathematical and Computer Modelling, Volume 46, Issues 1–2, July 2007, Pages 198–213, Proceedings of the International Conference on Computational Methods in Sciences and Engineering 2004". For UPSO parameters, for each folder described above, we have set some default values based on our experimentation on the various microarray/RNA-Seq datasets mentioned in the manuscript. We propose though to the user to experiment with parameters that may optimize the result like number of particles (too many can be highly time consuming), the neighborhood radius (increasing the neighbors can be highly time consuming) and the velocity. The space bounds have also been set based on experimentation on different datasets.

Note: To make it user-friendly, none of the solver-specific parameters are defined when calling the function. To do so, check the relevant functions and modify "options = optimset(….)." Similarly, modify all UPSO related functions.

6. CLUSTERING ALGORITHMS

Clustering algorithms such as k-means and k-medoids with distance 'correlation'. We experimented with all matlab distance functions and correlation performed best. The clustering of the data can be done on linear or log space although we note that the expression profiles used for deconvolution are always in linear space. Also, for kmeans the cluster exemplar is the average expression profile of the genes in the same cluster in linear space. For kmedoids, the cluster exemplar can be either the medoid (transformed in linear space if needed) or the average expression profile of the genes in the same cluster in linear space.

RUN EXAMPLES

'run_example_GSE19830_proportion_estimation.m': Test unsupervised mode of Deblender with GSE19830 dataset. It has 33 mixed samples including 3 tissues. In this example, all probes of the dataset are included except the control probes starting with AFFX-. The input is the mixed expression data 'mixed_data', the probe names 'mixed_genes'. The output is the proportion matrix 'A_estimated' and the ground truth proportion matrix is 'A_real'. The estimated proportions are plotted against true proportions.

'run_example_RNA_Seq_MDL_proportion_estimation.m': Test unsupervised mode of Deblender, i.e. calculating number of tissue types and mixture proportions, with RNA-Seq dataset downloaded by DECONRNASeq tool (Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics. 2013;29: 1083-5). The dataset has 10 mixed samples containing 5 different tissues. In the 'summarized_mixed_data' one RefSeq is chosen per gene name in case of multiple RefSeq Ids corresponding to the same gene name. An offset of 0.0001 was added to all values. The output is MDL and the proportion matrix 'A_estimated' and the ground truth proportion matrix is 'A_real'. MDL is plotted in the range {2,8} tissues and the estimated proportions are plotted against true proportions.

LICENCE

Author/Maintainer: Konstantina Dimitrakopoulou Konstantina.Dimitrakopoulou@uib.no, dimitrakopouloukon@gmail.com