

Core Mathematical Function of SlimR v1.0.8

Zhaoqing Wang

zhaoqingwang@mail.sdu.edu.cn

Section 1 Celltype Calculate

Variables

- C : Number of cell clusters from input cluster_col.
 - A : Number of cell clusters from the Markers_list.
 - N : Number of genes in specific cell types from the Markers_list.
 - $i \in \{1, C\}$: cluster_col's cluster index.
 - $a \in \{1, A\}$: Markers_list's cluster index.
 - $g \in \{1, N\}$: Markers_list's gene index corresponding to the current cell type.
 - $x_{g,i}$: Expression value of gene g in cluster i .
 - $\mu_{g,i}$: Average expression of gene g in cluster i .
 - $\sigma_{g,i}$: Standard deviation of gene g in cluster i .
 - $f_{g,i}$: Fraction of cells in cluster i where $x_{g,i} > m$ (minimum expression threshold m ; default: 0.1).
 - w : Specificity weight parameter (default: 3).
 - $\sigma_{g,i}$: Average standard deviation of all genes in cluster i .
 - ε : Small constant to avoid division by zero (default: 1×10^{-6}).
-

Step 1: Specificity Score Calculation

For each gene g in cluster i corresponding to the current cell type a in the Markers_list:

$$s_{g,i} = \mu_{g,i} \cdot f_{g,i} \cdot (1 + w \cdot \frac{\sigma_{g,i}}{\bar{\sigma}_i + \varepsilon})$$

Explanation:

- $\mu_{g,i}$: Mean expression level.
 - $f_{g,i}$: Proportion of cells expressing g .
 - $\sigma_{g,i}$: Normalized variability of g compared to other genes in the cluster.
 - w : Amplifies the impact of high variability.
-

Step 2: Normalization of Specificity Scores

Normalize $s_{g,i}$ across genes per cluster i :

$$s'_{g,i} = \frac{s_{g,i} - \min(s_{g,i})}{\max(s_{g,i}) - \min(s_{g,i})} \text{ (if } \max(s_{g,i}) \neq \min(s_{g,i}))$$

Purpose: Ensures scores are comparable across genes within the same cluster.

Step 3: Gene Weight Calculation

Compute weights for genes based on their variability-to-mean ratio:

$$g_w = \frac{sd(\sigma_{g,i})}{\text{mean}(\mu_{g,i})} \text{ (if } \text{mean}(\mu_{g,i}) \neq 0)$$

Purpose: Prioritize genes with higher variability and lower mean expression.

Step 4: Cluster-Specific Gene Expression Score

Aggregate normalized scores $s'_{g,i}$ into a final cluster score $p_{i,a}$ corresponding to the current cell type a in the Markers_list:

$$p_{i,a} = \sum_{g=1}^N g_w \cdot s'_{g,i}$$

Interpretation:

- $p_{i,a}$ reflects the weighted sum of gene-specificity scores for cluster i .
 - Higher $p_{i,a}$ indicates stronger evidence for the cluster corresponding to the current cell type a .
-

Step 5: Final Probability Matrix

For cluster_col's cluster $i \in \{1, C\}$ and Markers_list's clusters $a \in \{1, A\}$, the function outputs a probability matrix R where:

$$R_{i,a} = \frac{p_{i,a} - \min_{b \in A}(p_{i,b})}{\max_{b \in A}(p_{i,b}) - \min_{b \in A}(p_{i,b})} \quad (\text{if } \max_{b \in A}(p_{i,b}) \neq \min_{b \in A}(p_{i,b}))$$

Note: For each cluster_col's cluster i , the cell type with the highest normalized score p_i above threshold is selected as predicted cell type.

Step 6: AUC Validation

AUC Correction: ROC-AUC is computed using mean expression of signature genes to validate predictions:

$$AUC = \int_0^1 TPR(FPR) dFPR$$

Where:

$$TPR = f(\text{mean}(x_{g,i}))$$

Interpretation:

- True Positive Rate (TPR): $\text{Sensitivity} = \frac{TP}{TP + FN}$.
 - False Positive Rate (FPR): $1 - \text{Specificity} = \frac{FP}{FP + TN}$.
 - $x_{g,i}$: Expression value of gene g in cluster i .
-

Section 2 Celltype Verification

Variables

- $c \in \{1, A\}$: After annotation cluster index.
 - $g \in \{1, N\}$: Gene index.
 - $x_{g,i}$: Expression value of gene g in cluster i .
 - $f_{g,i}$: Fraction of cells in cluster i where $x_{g,i} > m$ (minimum expression threshold m ; default: 0.1).
 - k : Top gene count (default: 5).
-

Gene Scoring System

When the cell type c is in "Markers_list", verification markers uses the markers $g \in \{1, N\}$ corresponding to the specific cell type in it.

Screening of verification markers for cell types c not located in "Markers_list", compute each gene g in after annotation cell types c :

$$G_c^k = \tau_k \left(\sum_{c \neq j} \log_2 \left(\frac{\mu_{g,c}}{\bar{\mu}_{g,j}} \right) \cdot f_{g,c} \right)$$

Note: $\bar{\mu}_{g,j} = \text{mean}(\mu_{g,j} \text{ for all } j \neq c)$: Average mean expression in other clusters.

Feature Significance Score (FSS)

Feature Significance Score, FSS, product value of 'log2FC' and 'Expression ratio':

$$FSS = \Delta \log_2(\mu_{g,c}) \cdot f_{g,c}$$

Where:

$$\Delta \log_2(\mu_{g,c}) = \log_2 \left(\frac{\mu_{g,c}}{\bar{\mu}_{g,j}} \right)$$

Note: The 'FSS' parameter is also used in the 'Read_seurat_markers()' function for Markers screening.
