Core Mathematical Function of SlimR v1.0.7

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 types
- $x_{g,i}$: 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 types a :

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

Explanation:

- $\mu_{g,i}$: Mean expression level.
- $f_{g,i}$: Proportion of cells expressing g.
- $\sigma_{g,i}i$: 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 :

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

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:

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

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:

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

Interpretation:

- p_i reflects the weighted sum of gene-specificity scores for cluster i.
- Higher p_i indicates stronger evidence for the cluster corresponding to the current cell types 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:

$$R_{i,a} = \frac{p_{i,a} - \min\limits_{b \in A}(p_{i,b})}{\max\limits_{b \in A}(p_{i,b}) - \min\limits_{b \in A}(p_{i,b})} \ (if\ \max\limits_{b \in A}(p_{i,b}) \neq \min\limits_{b \in A}(p_{i,b}))$$

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

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：

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

Interpretation:

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

Section 2 Celltype Verification

Variables

- $c \in \{1, A\}$: After annotation cluster index
- $g \in \{1, \mathrm{N}\}$: 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 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:

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

Note: $\overline{\mu}_{g,j}$ = mean($\mu_{g,j}$ 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’:

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

Where:

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

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