rGSAn: An R package dedicated to the gene set analysis using semantic similarity measures

Aarón Ayllón Benítez, Fleur Mougin, Patricia Thébault
rGSAn: An R package dedicated to the gene set analysis using semantic similarity measures

Aarón Ayllón Benítez, Fleur Mougin, Patricia Thébault
Statistical enrichment analysis tool

Gene set

Reference genome or reference gene set

Annotation databases

Gene set annotation

Comparison using statistical methods

Reference annotation

List of enriched annotation terms

R tools

- gProfileR
  (Reimand et al. 2007)
- RDAVIDWebService
  (Fresno et al. 2013)
- Cluster profiler
  (Yu et al. 2012)
- topGO
  (Alexa et al. 2010)

LIMITATIONS

- Loss of information: Loss of annotated genes
- Redundant information that have to be minimized
Statistical enrichment analysis tool

Gene set → Gene set annotation

Gene set annotation → Comparison using statistical methods

Comparison using statistical methods → Reference annotation

Reference annotation → Annotation databases

Annotation databases → Gene set annotation

Gene set annotation → List of enriched annotation terms

R tools
- gProfileR (Reimand et al. 2007)
- RDAVIDWebService (Fresno et al. 2013)
- Cluster profiler (Yu et al. 2012)
- topGO (Alexa et al. 2010)

LIMITATIONS
- Loss of information: Loss of annotated genes
- Redundant information that have to be minimized
**Statistical enrichment analysis tool**

- Gene set
- Annotation databases
- Reference genome or reference gene set
- Gene set annotation
- Reference annotation
- Comparison using statistical methods
- List of enriched annotation terms

**R tools**
- gProfileR (Reimand et al. 2007)
- RDAVIDWebService (Fresno et al. 2013)
- Cluster profiler (Yu et al. 2012)
- topGO (Alexa et al. 2010)

**LIMITATIONS**
- Loss of information: Loss of annotated genes
- Redundant information that have to be minimized
The aim of this work

To interpret and analyze gene sets grouped according to a phenotype of interest.

As an alternative to enrichment statistical methods, we investigated:

- semantic similarity measures
- graph theory methods

The main objectives are to identify:

- the best compromise between the number of retained annotation terms that has to be drastically reduced
- the number of related genes that has to be as large as possible
- to propose original interactive visualization facilities dedicated to the multi-scale analysis of gene set annotations
Gene set

Automatic retrieval annotation information

![Gene Ontology logo]

<table>
<thead>
<tr>
<th></th>
<th>Term 1</th>
<th>Term 2</th>
<th>Term 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term 1</td>
<td>1</td>
<td>SS1</td>
<td>SS2</td>
</tr>
<tr>
<td>Term 2</td>
<td>SS1</td>
<td>1</td>
<td>SS3</td>
</tr>
<tr>
<td>Term 3</td>
<td>SS2</td>
<td>SS3</td>
<td>1</td>
</tr>
</tbody>
</table>
Gene set

Automatic retrieval annotation information

Table:

<table>
<thead>
<tr>
<th>Term</th>
<th>Term 1</th>
<th>Term 2</th>
<th>Term 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term 1</td>
<td>1</td>
<td>SS1</td>
<td>SS2</td>
</tr>
<tr>
<td>Term 2</td>
<td>SS1</td>
<td>1</td>
<td>SS3</td>
</tr>
<tr>
<td>Term 3</td>
<td>SS2</td>
<td>SS3</td>
<td>1</td>
</tr>
</tbody>
</table>
Gene set

Automatic retrieval annotation information

Gene Ontology

Table:

<table>
<thead>
<tr>
<th>Term</th>
<th>Term 2</th>
<th>Term 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term 1</td>
<td>SS1</td>
<td>SS2</td>
</tr>
<tr>
<td>Term 2</td>
<td>SS1</td>
<td>1</td>
</tr>
<tr>
<td>Term 3</td>
<td>SS2</td>
<td>SS3</td>
</tr>
</tbody>
</table>

Diagram:

1. Gene set
2. Table
3. Diagram with labels R1, R2, R3, R4
4. Diagram with labels a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z
Automatic retrieval annotation information

Gene set

<table>
<thead>
<tr>
<th>Term 1</th>
<th>Term 2</th>
<th>Term 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SS1</td>
<td>SS2</td>
</tr>
<tr>
<td>SS1</td>
<td>1</td>
<td>SS3</td>
</tr>
<tr>
<td>SS2</td>
<td>SS3</td>
<td>1</td>
</tr>
</tbody>
</table>

Automatic retrieval annotation information

R1
R2
R3
R4

3
Automatic retrieval annotation information

Gene set

<table>
<thead>
<tr>
<th>Term 1</th>
<th>Term 2</th>
<th>Term 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SS1</td>
<td>SS2</td>
</tr>
<tr>
<td>SS1</td>
<td>1</td>
<td>SS3</td>
</tr>
<tr>
<td>SS2</td>
<td>SS3</td>
<td>1</td>
</tr>
</tbody>
</table>

R1 R2 R3 R4
- **DAG → TREE**
- Combination of two tree visualization metaphors
  - Circular tree map
  - Indented tree
- Interactive facilities to explore the results
rGSAn:

- R package
- With native functions in C++
- Import ontology (in OBO format)
- Insert annotation
  - From Go3AnnDbBimap R Class
  - From GAF file using flavin package
- Run the GSAn method

Output:

- Annotation results in a data frame for further purpose
  - Export files in CSV format
  - Interactive visualization using htmltools
**rGSAn:**

- R package
- With native functions in C++
- Import ontology (in OBO format)
- Insert annotation
  - From `Go3AnnDbBimap` R Class
  - From GAF file using `flavin` package
- Run the GSAn method

**Output:**

- Annotation results in a data frame for further purpose
  - Export files in CSV format
  - Interactive visualization using `htmltools`

**Thanks**

**Git repository:**
[https://bitbucket.org/Ayllonbe/rgsan](https://bitbucket.org/Ayllonbe/rgsan)

**Contact:**
aaron.aylon-benitez@u-bordeaux.fr