ViSEAGO: Easier data mining of biological functions organized into clusters using Gene Ontology and semantic similarity

Aurélien Brionne, Amélie Juanchich, Christelle Hennequet-Antier

ViSEAGO: Easier data mining of biological functions organized into clusters using Gene Ontology and semantic similarity

BOA, INRA, Université de Tours, 37380 Nouzilly, France

Aurélien Brionne: aurelien.brionne@inra.fr

Amélie Juanchich: amelie.juanchich@inra.fr

Christelle Hennequet-Antier: christelle.hennequet-antier@inra.fr

ViSEAGO

Visualization, **S**emantic similarity and **E**nrichment **A**nalysis of **G**ene **O**ntology.

R package publicly available on https://forgemia.inra.fr/umr-boa/viseago.

Objective:

Data mining of biological functions and establish links between genes

ViSEAGO:

→ Allows complex experimental design (multiple comparisons, large datasets)

→ Extends classical functional GO analysis to focus on functional coherence

→ Provides both a synthetic and detailed view using interactive functionalities respecting the GO graph structure.

State of the art

Comparison of different tools focused on biological interpretation from GO annotation

Tool	Enrichment test	GO terms SS	Sets of GO terms SS	Visualization (focus)	Multiple lists	Graph interactivity
David https://david.ncifcrf.gov/	Fisher Exact (EASE)	No	No	genes	Yes	No
ClusterProfiler Bioconductor	Hypergeometric	IC-based, Graph-based	Yes	genes	Yes	No
gProfiler https://biit.cs.ut.ee/gprofiler/	Hypergeometric	No	No	genes	Yes	No
REVIGO http://revigo.irb.hr/	No	IC-based	No	GO terms	No	Yes
ViSEAGO https://forgemia.inr.fr/umr-boa/viseago	Fisher Exact, Hypergeometric	IC-based, Graph-based	Yes	GO terms	Yes	Yes

Pipeline

GO Annotation of list(s) of features

Annotation database (Ensembl, NCBI, Uniport…)

Enrichment **A**nalysis

TopGO (Fisher-exact test)

	∈ GO term	∉ GO term
∈ list of interest	X	X
∉ list of interest	X	X

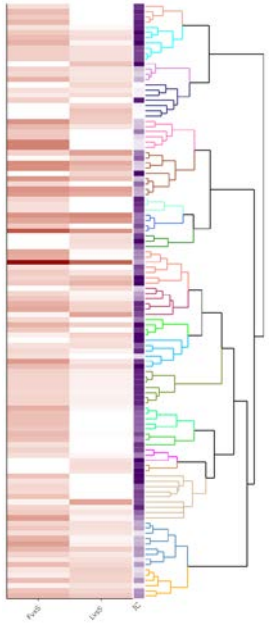
Semantic Similarity & **Vi**sualization

Semantic similarity: GOSemSim

Clustering: hclust, dynamicTreeCut

Interactivity: plotly

Wang GO terms distance clustering heatmap plot

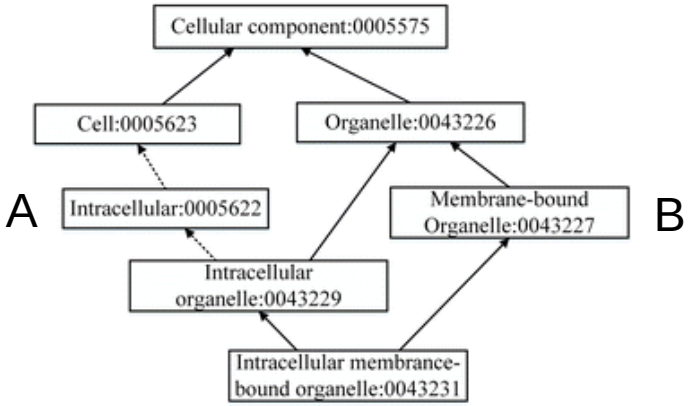


Biological Interpretation

Method: Wang's semantic similarity

Semantic similarity between 2 GO terms

$$S_{GO}(A, B) = ?$$



A

GO terms	A=0005622	0005623	**0005575**
Svalue	1	0.6	**0.48**

B

GO terms	B=0043227	0043226	**0005575**
Svalue	1	0.8	**0.64**

S-value of ancestor GO term (t) related to term A

contribution of term t to the semantic of term A

$$S_A(A) = 1$$

$$S_A(t) = max\ \{w_e\ x\ S_A(t') \mid t' \in \text{children of } (t)\ \} \text{ if } t \neq A$$

Semantic Value of GO term A, and B

$$SV(A) = \sum_{t \in T_A} S_A(t)$$

SV(A)=1+0.6+0.48=2.08

SV(B)=1+0.8+0.64=2.44

Semantic similarity between GO terms A and B

$$S_{GO}(A,B) = \frac{\sum_{t \in T_A \cap T_B} (S_A(t) + S_B(t))}{SV(A) + SV(B)}$$

$$\sum_{t \in TA \cap TB} (SA(t) + (S_B(t)) = SA(\mathbf{0005575}) + SB(\mathbf{0005575}) = 0.48 + 0.64$$

$$S_{GO}(A,B) = \frac{1.12}{2.44 + 2.08} = 0.25$$

Tian Z., *et al.*, BMC Bioinformatics, 2016

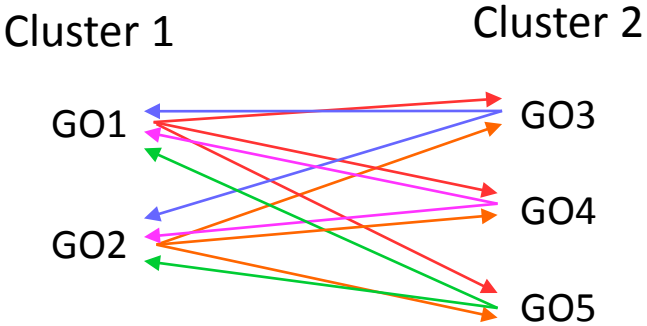
Method: BMA semantic similarity

Semantic similarity between 2 sets of GO terms

Best Match Average (BMA)

$$sim_{BMA}(g_1, g_2) = \frac{\sum_{1=i}^{m} \max_{1 \le j \le n} sim(go_{1i}, go_{2j}) + \sum_{1=j}^{n} \max_{1 \le i \le m} sim(go_{1i}, go_{2j})}{m+n}$$

→ average of all maximum similarities over all pairs of GO terms between two GO term sets



Wang's similarity between two GO terms

		Cluster 1		Cluster 2		
Cluster 1	GO1	x	x	0.3	0.3	**0.4**
	GO2	x	x	0.2	0.4	**0.5**
Cluster 2	GO3	**0.3**	0.2	x	x	x
	GO4	0.3	**0.4**	x	x	x
	GO5	0.4	**0.5**	x	x	x

$$\mathrm{sim}_{\mathrm{BMA}}(\mathrm{cluster1,cluster2}) = \frac{(0.4+0.5)+(0.3+0.4+0.5)}{2+3} = 0.75$$

Example: Hypomethylation in bull sperm targets specific genomic functions

hypomethylated CpGs genomic regions (HR) and their associated functions from MeDIP datasets in bull sperm in comparison to bovine somatic cells (fibroblasts and liver cells)

→ Re-used and re-analyzed Methylated DNA immunoprecipitation (MeDIP) dataset (GSE102960, Perrier J-P, *et al.*, BMC Genomics, 2018).

Gene Sets:

bull sperm in comparison to somatic cells: fibroblast (FvsS) and Liver (LvsS).

Genes from HR with match in regulatory elements → FvsS: 1632 HR, **732** genes
LvsS: 3109 HR, **1229** genes

promoter: -1kb to +0.1 kb along TSS;
downstream: +1kb along TES

background: 24 616 genes

GO annotation and enrichment analysis

Lists of available genes
732 FvsS genes
1229 LvsS genes

24 616 background genes

Annotation database
btaurus Ensembl version 81
Ensembl2GO()
annotate()

Lists of feasible genes
(with at least one GO annotation)
495 FvsS genes
788 LvsS genes

15040 background genes

Enrichment Analysis
topGO package, Fisher's exact test, Biological Process (BP)
create_topGOdata(),
RunTest(),
merge_enrich_terms()

Enriched GO terms (p<0.01)
91 FvsS
53 LvsS

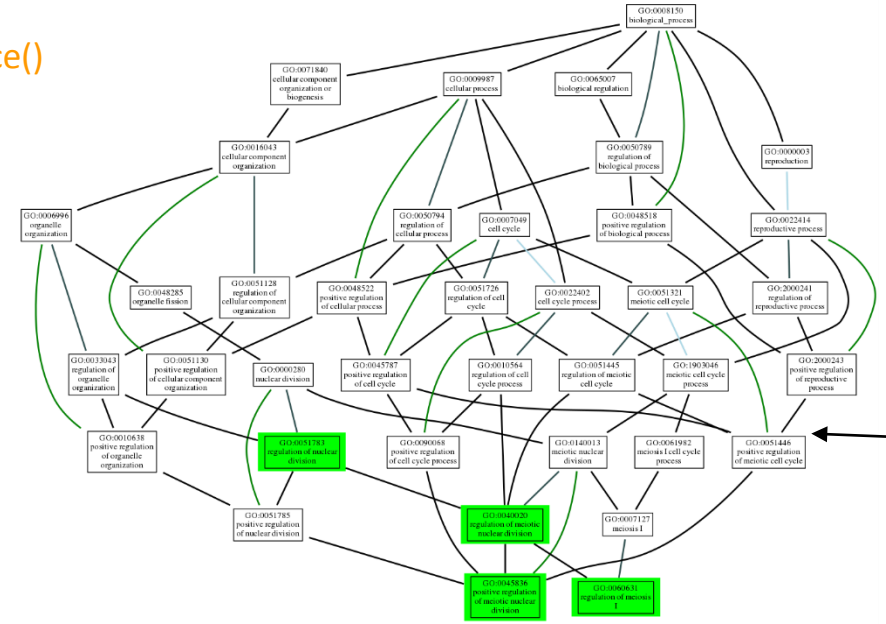
114 in at least one comparison (shared GO terms: 30)

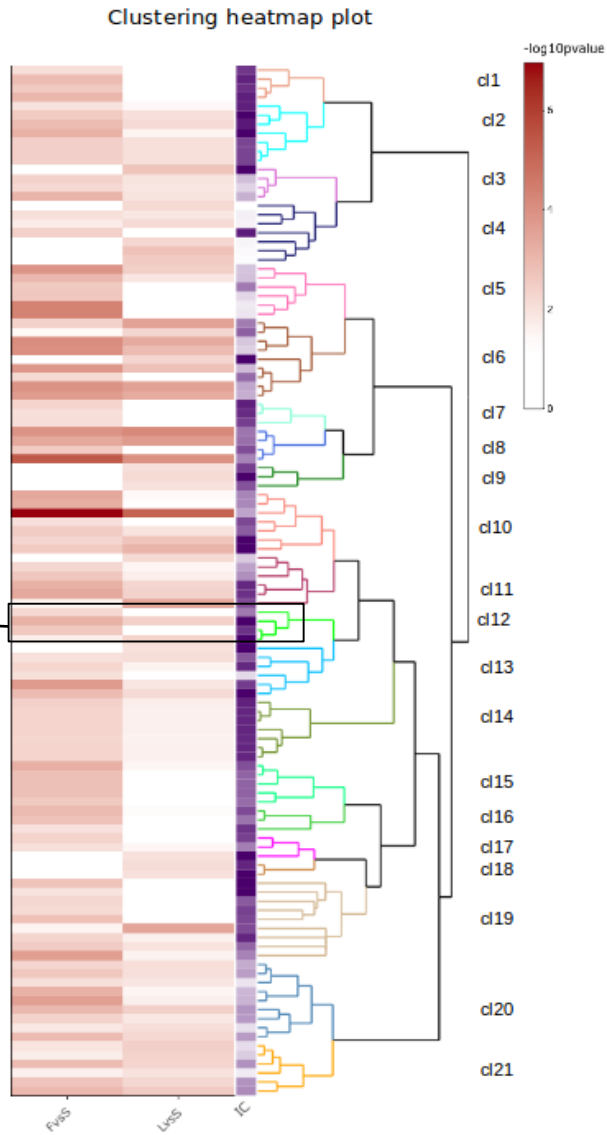
Organized enriched GO terms

hierarchical clustering on GO terms

Wang's method, *ward.D2* aggregation criterion, dynamically cut

build_GO_SS()
compute_SS_distance()
GOterms_heatmap()
show_heatmap()





Cluster 12 → 4 GO terms

13 genes involved in regulation of nuclear division (meiosis)

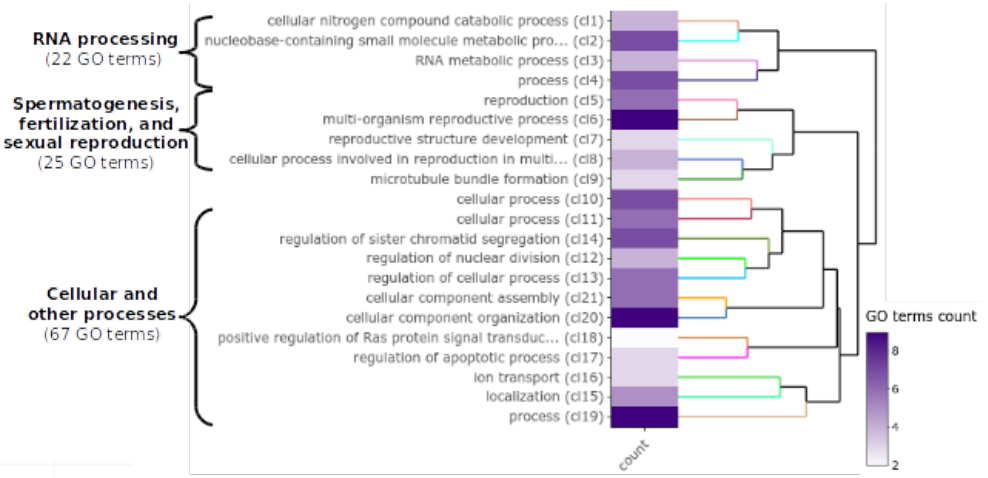
MAP9, PIWIL2, ENSBTAG00000005708, LCMT1, PRDM9, ENSBTAG00000024874, ENSBTAG00000035129, ENSBTAG00000035319, DAZL, CALR, UBE2C, MSX1, UBE2B

Organized sets of enriched GO terms

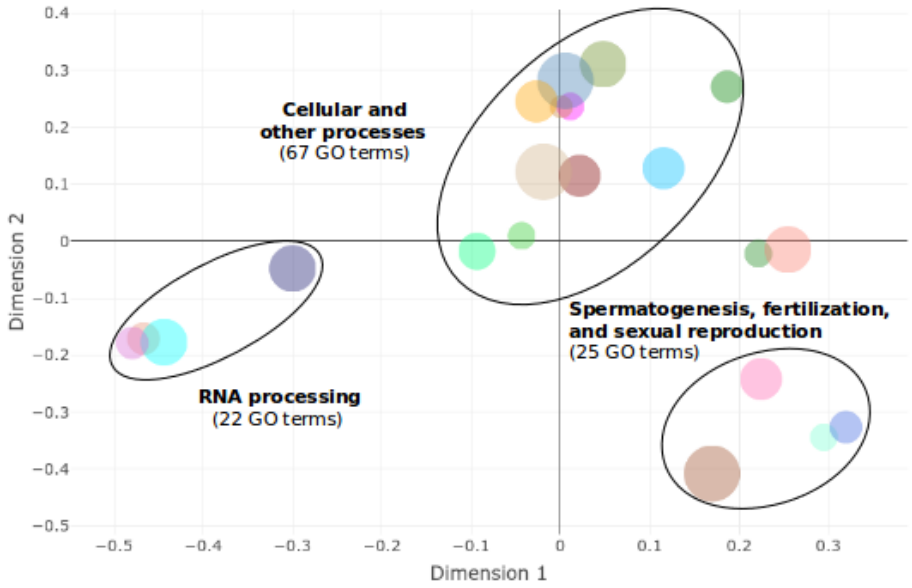
Heatmap and MDSplot on sets of GO terms
BMA method, *ward.D2* aggregation criterion

compute_SS_distance()
GOclusters_heatmap()
show_heatmap()
MDSplot()

heatmap plot of functional sets of GO terms



MDS plot of functional sets of GO terms



Conclusion

ViSEAGO: **data mining of biological functions using GO terms**

→ Semantic similarity and visualization

***ViSEAGO*'s functionalities:**

(1) emphasize functional coherence

(2) reliability of the functional interpretation

(3) facilitate biological interpretation

interactive visualization both synthetic and detailed

***ViSEAGO* helps users to perform a reproducible functional analysis and to prioritize genes to investigate.**

Thank you for your attention

Thanks to UMR BOA
Thanks to Perrier J-P, *et al.* (BMC Genomics, 2018) for GSE102960 MeDIP dataset

R package publicly available on https://forgemia.inra.fr/umr-boa/viseago