HAL open science

Development of new computational methods for a synthetic gene set annotation

Aarón Ayllón-Benítez

To cite this version:

Aarón Ayllón-Benítez. Development of new computational methods for a synthetic gene set annotation. Bioinformatics [q-bio.QM]. Université de Bordeaux, 2019. English. NNT : 2019BORD0305 . tel-02401947v2

HAL Id: tel-02401947

https://hal.science/tel-02401947v2

Submitted on 2 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LABORATOIRE BORDELAIS DE RECHERCHE EN INFORMATIQUE

École doctorale Mathématiques et informatique

université de BORDEAUX

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

par **Aarón Ayllón Benítez**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

Development of new computational methods for a synthetic gene set annotation

---

Développement de nouvelles méthodes informatiques pour une annotation synthétique d'un ensemble de gènes

---

Date de soutenance : 5 décembre 2019

Devant la commission d'examen composée de:

Olivier Dameron	Maître de Conférences HDR, Université de Rennes 1	Rapporteur
Mohamed Elati	Professeur, Université de Lille	Rapporteur
Antoine de Daruvar	Professeur, Université de Bordeaux	Président
Christine Gaspin	Directrice de Recherche, INRA Toulouse	Examinatrice
Jesualdo Tomás Fernández Breis	Professeur, Universidad Murcia (Espagne)	Examinateur
Patricia Thébault	Maître de Conférences HDR, Université de Bordeaux	Directrice
Fleur Mougin	Maître de Conférences HDR, Université de Bordeaux	Co-directrice

– 2019 –

Abstract

Development of new computational methods for a synthetic gene set annotation

by Aarón Ayllón Benítez

The revolution in new sequencing technologies, by strongly improving the production of omics data, is greatly leading to new understandings of the relations between genotype and phenotype. To interpret and analyze data grouped according to a phenotype of interest, methods based on statistical enrichment became a standard in biology. However, these methods synthesize the biological information by *a priori* selecting the over-represented terms and focus on the most studied genes that may represent a limited coverage of annotated genes within a gene set. During this thesis, we explored different methods for annotating gene sets. In this frame, we developed three studies allowing the annotation of gene sets and thus improving the understanding of their biological context.

First, visualization approaches were applied to represent annotation results provided by enrichment analysis for a gene set or a repertoire of gene sets. In this work, a visualization prototype called `MOTVIS` (*MOdular Term VISualization*) has been developed to provide an interactive representation of a repertoire of gene sets combining two visual metaphors: a treemap view that provides an overview and also displays detailed information about gene sets, and an indented tree view that can be used to focus on the annotation terms of interest. `MOTVIS` has the advantage to solve the limitations of each visual metaphor when used individually. This illustrates the interest of using different visual metaphors to facilitate the comprehension of biological results by representing complex data.

Secondly, to address the issues of enrichment analysis, a new method for analyzing the impact of using different semantic similarity measures on gene set annotation was proposed. To evaluate the impact of each measure, two relevant criteria were considered for characterizing a "good" synthetic gene set annotation: (i) the number of annotation terms has to be drastically reduced while maintaining a sufficient level of details, and (ii) the number of genes described by the selected terms should be as large as possible. Thus, nine semantic similarity measures were analyzed to identify the best possible compromise between both criteria while maintaining a sufficient level of details. Using *Gene Ontology* (GO) to annotate the gene sets, we observed better results with *node-based* measures that use the terms' characteristics than with *edge-based* measures that use the relations terms. The annotation of the gene sets achieved with the *node-based* measures did not exhibit major differences regardless of the characteristics of the terms used. Then, we developed `GSAn` (*Gene Set Annotation*), a novel gene set annotation web server that uses semantic similarity measures to synthesize *a priori* GO annotation terms. `GSAn` contains the interactive visualization `MOTVIS`, dedicated to visualize the representative terms of gene set annotations. Compared to enrichment analysis tools, `GSAn` has shown excellent results in terms of maximizing the gene coverage while minimizing the number of terms.

At last, the third work consisted in enriching the annotation results provided by `GSAn`. Since the knowledge described in GO may not be sufficient for interpreting gene sets, other biological information, such as pathways and diseases, may be useful to provide a wider biological context. Thus, two additional knowledge resources, being *Reactome* and *Disease Ontology* (DO), were integrated within `GSAn`. In practice, GO terms were mapped to terms of *Reactome* and DO, before and after applying the `GSAn` method. The integration of these resources improved the results in terms of gene coverage without affecting significantly the number of involved terms. Two strategies were applied to find mappings (generated or extracted from the web) between each new resource and GO. We have shown that a mapping process before computing the `GSAn` method allowed to obtain a larger number of inter-relations between the two knowledge resources.

Résumé

Développement de nouvelles méthodes informatiques pour une annotation synthétique d'un ensemble de gènes

by Aarón Ayllón Benítez

Les avancées dans l'analyse de l'expression différentielle de gènes ont suscité un vif intérêt pour l'étude d'ensembles de gènes présentant une similarité d'expression au cours d'une même condition expérimentale. Les approches classiques pour interpréter l'information biologique reposent sur l'utilisation de méthodes statistiques. Cependant, ces méthodes se focalisent sur les gènes les plus connus tout en générant des informations redondantes qui peuvent être éliminées en prenant en compte la structure des ressources de connaissances qui fournissent l'annotation. Au cours de cette thèse, nous avons exploré différentes méthodes permettant l'annotation d'ensembles de gènes. Dans ce cadre, nous avons élaboré trois travaux permettant l'annotation d'ensembles de gènes pour améliorer la compréhension de leur contexte biologique.

Premièrement, nous présentons les solutions visuelles développées pour faciliter l'interprétation des résultats d'annota-tion d'un ou plusieurs ensembles de gènes. Dans ce travail, nous avons développé un prototype de visualisation, appelé `MOTVIS`, qui explore l'annotation d'une collection d'ensembles des gènes. `MOTVIS` utilise ainsi une combinaison de deux vues inter-connectées : une arborescence qui fournit un aperçu global des données mais aussi des informations détaillées sur les ensembles de gènes, et une visualisation qui permet de se concentrer sur les termes d'annotation d'intérêt. La combinaison de ces deux visualisations a l'avantage de faciliter la compréhension des résultats biologiques lorsque des données complexes sont représentées.

Deuxièmement, nous abordons les limitations des approches d'enrichissement statistique en proposant une méthode originale qui analyse l'impact d'utiliser différentes mesures de similarité sémantique pour annoter les ensembles de gènes. Pour évaluer l'impact de chaque mesure, nous avons considéré deux critères comme étant pertinents pour évaluer une annotation synthétique de qualité d'un ensemble de gènes : (i) le nombre de termes d'annotation doit être réduit considérablement tout en gardant un niveau suffisant de détail, et (ii) le nombre de gènes décrits par les termes sélectionnés doit être maximisé. Ainsi, neuf mesures de similarité sémantique ont été analysées pour trouver le meilleur compromis possible entre réduire le nombre de termes et maintenir un niveau suffisant de détails fournis par les termes choisis. Tout en utilisant la *Gene Ontology* (GO) pour annoter les ensembles de gènes, nous avons obtenu de meilleurs résultats pour les mesures de similarité sémantique basées sur les nœuds qui utilisent les attributs des termes, par rapport aux mesures basées sur les arêtes qui utilisent les relations qui connectent les termes. Enfin, nous avons développé `GSAn`, un serveur web basé sur les développements précédents et dédié à l'annotation d'un ensemble de gènes *a priori*. `GSAn` intègre `MOTVIS` comme outil de visualisation pour présenter conjointement les termes représentatifs et les gènes de l'ensemble étudié. Nous avons comparé `GSAn` avec des outils d'enrichissement et avons montré que les résultats de `GSAn` constituent un bon compromis pour maximiser la couverture de gènes tout en minimisant le nombre de termes.

Le dernier point exploré est une étape visant à étudier la faisabilité d'intégrer d'autres ressources dans `GSAn` pour améliorer les résultats. Nous avons ainsi intégré deux ressources, l'une décrivant les maladies humaines avec *Disease Ontology* (DO) et l'autre les voies métaboliques avec *Reactome*. Le but était de fournir de l'information supplémentaire aux utilisateurs finaux de `GSAn`. Nous avons évalué l'impact de l'ajout de ces ressources de connaissances dans `GSAn` lors de l'analyse d'ensembles de gènes. L'intégration de ces ressources a amélioré les résultats en couvrant d'avantage de gènes sans pour autant affecter de manière significative le nombre de termes impliqués. Ensuite, les termes GO ont été mis en correspondance avec les termes DO et *Reactome*, *a priori* et *a posteriori* des calculs effectués par `GSAn`. Nous avons montré qu'un processus de mise en correspondance appliqué *a priori* permettait d'obtenir un plus grand nombre d'inter-relations entre les deux ressources.

Acknowledgements

Je tiens tout d'abord à remercier vivement Olivier Dameron et Mohamed Elati d'avoir accepté d'être rapporteurs de mon travail malgré leur emploi du temps très chargé. Je suis très heureux de bénéficier de leur avis sur mes travaux de recherche et j'aimerais que cela soit l'occasion de mettre en place des collaborations futures.

Je remercie Antoine de Daruvar, Christine Gaspin et Jesualdo Tomás Fernández Breis de participer à mon jury. Je voudrais aussi remercier Jesualdo de m'avoir initié au monde de la bio-informatique. Estoy seguro que gracias a ti fui seleccionado para hacer las practicas que han llevado mas tarde a esta tesis.

Je tiens à remercier Patricia Thébault et Fleur Mougin qui m'ont dirigé et m'ont offert l'opportunité de faire une thèse. Au début ce n'était pas facile, surtout pour le français qui a demandé beaucoup d'investissement de votre part. Je reconnais que mes explications durant certains rendez vous tout au long de ce thèse n'étaient pas toujours évidentes, surtout pour les réunions en fin après midi. Vous étiez toujours disponible pour moi et m'avez permis (surtout Patricia par proximité de bureau) de vous embêter avec des réunions non programmées. Au cours de ces trois (et quelques) années, j'ai beaucoup appris de vous deux et j'espère un jour pouvoir arriver à votre niveau. Individuellement vous avez un très grand niveau, autant personnel que professionnel, mais vous deux ensemble, la combinaison était parfaite. Ce fut un réel plaisir que de travailler avec vous, et j'espère que sincèrement que cela pourra continuer.

Merci à chaque intégrant du bureau Jonh Snow (ISPED). Je voudrais remercier Jean Noël, Irène, Solène, Marie et Laura. Avec vous, le journées où j'étais dans ce bureau se sont bien passées ainsi que les milliers de soirées ensembles que je gardes dans mon coeur. Aussi, dans le même bureau je voudrais spécialement remercier deux personnes. En premier lieu, Hadrien, qui a été une merveilleuse découverte (je n'ai jamais rencontré une personne qui fait autant de conneries que lui et moi). Hadrien est pour moi un mathématicien d'excellence ainsi qu'un pédagogue, mais surtout un très bon ami. Avec son tic de langage : "Aarón il faut absolument que tu fasses un cours de mats", Hadrien finit toujours par m'expliquer ce dont j'ai besoin. En deuxième lieu, mais ça n'est pas moins important, je voudrais remercier très fortement Bruno ("la machine sauvage"). Il a toujours été là pour tout. Mon compagnon de sport, même si on a pas de tablette, j'ai passé des moments merveilleux avec toi.

Évidement, je n'oublie pas d'autres ispediens (et ispediennes) comme c'est le cas d'Anaïs, Boris, Robin, Corentin, Soufiane, Sophie, Camille, Maude, Chloé, Emilie, Perrine, Cinciana *et al.* Je n'oublie pas non plus les membres de l'équipe Erias, qui, même si mon domaine d'expertise était diffèrent, m'ont toujours bien accueilli (surtout Vianney qui me considère un très bon doctorant après Jean Noël). Je remercie Rodolphe Thiébaut pour son support et ainsi donner l'opportunité de collaborer avec lui pendant la thèse.

À 3.7 Km de distance de l'ISPED (et ~10 minutes à velo), je voudrais remercier à mes collègues du LaBRI (d'aujourd'hui et d'hier). Je remercie profondément Romain Bourqui. Il m'a beaucoup appris et ouvert (partiellement) les portes du monde de la visualisation de l'information. De plus, je le remercie pour toutes les diverses questions administratives dans le monde de la recherche. Gaëlle et Frederic (collègues de bureau 306) qui m'ont toujours aidé pour des problèmes techniques et rigolé aux blagues très nulles. Aussi Gaëlle, qui a commencé au même moment que moi, pour le soutien mutuel, les longues soirées de travail au bureau pour rentrer ainsi ensemble (à vélo), ainsi que les soirées jeux et le cinéma ensemble. Aux 4 docteurs qui sont partis pour refaire leur vie: Alexandre, Claire, Joris et Antoine.

Je remercie aussi mes deux beaux gosses, Jason et Norbert à qui je souhaite une belle continuation avec *IntuiNet*. L'ensemble de l'équipe Bkb (ancien MaBioVis) ainsi que les membres du conseil de laboratoire pour avoir écouté mes suggestions (même si il n'y a pas de machine sèche main Dyson dans les toilettes pour ne plus utiliser le papier). Aussi, je n'oublie pas l'équipe d'administration qui m'a toujours aidé avec un grand sourire, spécialement Isabelle, Emanuela et Cathy.

Merci infiniment à Marthe, pour m'avoir laissé squatter chez elle pendant mes derniers mois de thèse. Tu m'as vraiment aidé dans un moment difficile pour moi. Aussi, Maylis (coloc de Marthe), qui a été une surprise très agréable et que j'espère que ça durera dans le temps.

Aussi bien sur, merci à Anna : je n'oublierai jamais les longues discussions, la découverte de la bière tchèque et les conneries quand on danse. Aussi, te remercier de ne pas avoir eu de remords quand je t'ai promis de faire une soirée broderie que je n'ai jamais faite.

Pour finir, je voudrais remercier ma famille et mes amis d'Espagne. En primer lugar, los Caravaca, a los que agradezco por todo. Gracias a Manuel y Jaime por las risas y las tonterías y sobre todo a su amistad incondicional.

Me gustaría agradecer también a Pepe, por las grandes conversaciones donde no sacamos nada en claro y por las partidas a la consola, donde claro está, siempre gano yo.

A toda mi familia, gracias. A mi tita María José y mi tito Carlos, gracias por ser tan buenos y darme tantas alegrías. A mis hermanos, que aunque hay días que los quiero matar (literalmente), los quiero. A mi madre, porque durante toda la vida ha creído en mí, ha estado conmigo en mis depresiones y ha confiado en mí para seguir adelante y llegar donde he llegado. Por supuesto, a mi padre, que me ha enseñado muchísimas cosas en la vida y que es una pena que no siga con nosotros para que pueda seguir enseñándome. Siempre estará en mi corazón.

Contents

Abstract iii

Résumé v

Acknowledgements vii

Research contributions 1

Résumé substantiel 3

1 Introduction 13
- 1.1 The role of bioinformatics into biological and clinical researches . . . 13
- 1.2 Motivation to improve the information ciphered by gene sets . . . 14
- 1.3 Project outline . . . 15
- 1.4 Datasets used in this thesis . . . 16

2 Knowledge resources and functional annotation 19
- 2.1 Information supplied by a gene . . . 19
- 2.2 Knowledge resources providing biological annotation . . . 20
  - 2.2.1 Gene Ontology . . . 21
  - 2.2.2 Other knowledge resources . . . 26
  - 2.2.3 Limitation in studying single genes . . . 27
- 2.3 Interpreting gene sets . . . 28
  - 2.3.1 Gene set annotation by classical enrichment techniques . . . 29
  - 2.3.2 Opportunities of ontological annotation . . . 31
  - 2.3.3 Summarizing results with visualization . . . 34
  - 2.3.4 Using multiple knowledge resources to improve the biological understanding . . . 37
- 2.4 Challenges related to the annotation of gene sets . . . 39

3 Gene sets visualization 43
- 3.1 Visualizing the annotation of a single gene set . . . 44
  - 3.1.1 Computation of the gene set annotation . . . 45
  - 3.1.2 Visual metaphors to represent the annotation of a single gene set . . . 46
  - 3.1.3 Case study . . . 49
- 3.2 Visualizing the annotation of multiple gene sets . . . 50
  - 3.2.1 Annotating gene sets . . . 53
  - 3.2.2 Relating annotation terms to GO terms . . . 53
  - 3.2.3 GO structure simplification . . . 53
  - 3.2.4 Visualizing with `MOTVIS` . . . 54
  - 3.2.5 Case study . . . 56
- 3.3 Conclusion . . . 58

4 Annotating gene sets by using semantic similarity measures **61**
4.1 Gene set annotation with semantic similarity measures 62
4.1.1 State of the art of alternatives to enrichment analysis 62
4.1.2 Motivation to develop a method based on semantic similarity measures for annotating gene sets 62
4.2 Impact of semantic similarity in the gene set annotation 63
4.2.1 Semantic similarity measures 64
4.2.2 Selection of Gene Ontology terms 67
4.2.3 Clustering annotation terms using computed semantic similarity matrices 68
4.2.4 Identifying the most relevant representative terms 74
4.2.5 Discussion 80
4.3 The extension of the analysis framework 81
4.3.1 The GSAn method 82
4.3.2 The GSAn web server 84
4.3.3 Comparison of GSAn with enrichment analysis tools 89
4.3.4 Case study 93
4.3.5 GSAn in the R language or RGSAn 94
4.4 Conclusions 95

5 Integration of additional knowledge resources within GSAn **97**
5.1 Existing computational solutions that integrate knowledge resources 98
5.2 Description of knowledge resources to be included within GSAn 99
5.3 Pre-process of resources before their integration within the GSAn framework . 99
5.3.1 Adapting the structure of *Disease Ontology* and *Reactome* 100
5.3.2 Recovering annotations from DO and *Reactome* 101
5.4 Evaluating the addition of a knowledge resource in GSAn without any integration 102
5.5 Mapping methods to align new knowledge resources to GO 105
5.5.1 Mapping Disease Ontology terms to Gene Ontology terms 106
5.5.2 Mapping *Reactome* terms to GO terms 107
5.6 Comparison between an *a priori* and *a posteriori* integration 108
5.6.1 Integrating DO within GSAn 108
5.6.2 Integrating *Reactome* within GSAn 110
5.7 Conclusion 111

6 Conclusions and research perspectives **113**
6.1 Review and findings 113
6.2 Research perspectives 115
6.3 Final conclusion 119

Bibliography **121**

A Additional information for chapter 4 **149**

B Additional information for chapter 5 **151**

List of Figures

1.1 Decreasing tendency in (A) megabase DNA sequences' cost and (B) genome cost from September 2001 to February 2019. 14

2.1 Simplified vision of the central dogma of molecular biology 20
2.2 Screenshot from `Protégé` [Noy+03] showing information related to the GO term *cytolysis* 22
2.3 GO excerpt from the `QuickGO` website [Bin+09b] 23
2.4 *Gene Association File* (GAF) 2.1 column description 26
2.5 Enrichment analysis schemas for the three classes: (A) SEA or ORA, (B) GSEA or FCS, and (C) MEA and its sub-category PT 30
2.6 Examples of visualization results proposed by different enrichment tools . . . 36

3.1 Implemented pipeline to annotate and visualize a single gene set 45
3.2 (A) A schema representing the creation process of the treemap according to Johnson and Shneiderman [JS91]. (B) The possible interactions proposed in the visualization 47
3.3 Visual description of the loom layout 48
3.4 Results of the annotation of the gene set annotated as *Interferon* by Chaussabel and Baldwin [CB14] 49
3.5 Proposed visualizations of the annotation of a gene set 51
3.6 Pipeline to annotate and visualize multiple gene sets 52
3.7 Illustration of the process of GO simplification 54
3.8 Indented tree and treemap view representing the immune response within the [`C-260`] dataset 55
3.9 Zoom on the gene sets whose annotation provided by the enrichment analysis has been completely or partially mapped to the GO term *signaling* when `OntoEnrich` was applied 57
3.10 `MOTVIS` view focused on the M1.2 gene set. 58

4.1 Proposed workflow to annotate gene sets and to study the impact of using different semantic similarity measures 63
4.2 Classification of the nine semantic similarity measures investigated in this study according to the features that they use (adaptation from Guzzi et al. [Guz+12]) . 66
4.3 Depth of GO terms in the annotations provided by GOA human 68
4.4 Cophenetic correlation coefficients (*CCC*) according to nine semantic similarity measures of the investigated gene sets 71
4.5 Average Silhouette Width (*ASW*) according to nine semantic similarity measures of the investigated gene sets 72
4.6 *Z-index* scores for comparing the whole structure of the dendrograms using CLHM and ALHM for the investigated gene sets 73
4.7 Steps for identifying the representative terms given a cluster of GO terms . . . 75
4.8 IC_{GOu} distribution of GO terms used in GOA human 77

4.9 Number of representative terms obtained after applying each semantic similarity measure for the investigated gene sets 78
4.10 Percentage of synthetic terms using each semantic similarity measure and comparison with the DAVID enrichment tool 78
4.11 Percentage of covered genes using each semantic similarity measure and comparison with the DAVID enrichment tool 79
4.12 Comparison between the method proposed to annotate gene sets in section 4.2 and the method implemented in GSAn 81
4.13 Example of *Synthetic Algorithm* (*SA*) with six representative terms 83
4.14 GSAn output results (1) 87
4.15 GSAn output results (2) 87
4.16 GSAn output results (3) 88
4.17 Box-plots providing the impact on the number of terms for each tool 90
4.18 Box-plots providing the impact on the gene coverage for each tool 91
4.19 Box-plots showing the classification of terms by 2, 3, 4, 5 and more than 5 annotated genes 92
4.20 Report on the number of GSAn's users in the world from November 18th 2018 (start date of the server) to September 29th 2019 94

5.1 Data storage details of each organism within the *Reactome* release version 69. . 100
5.2 Number of (A) gene-DO term associations and (B) gene-*Reactome* term associations. 103
5.3 Number of representative terms computed by GSAn using DO, *Reactome* and GO independently as well as DO+*Reactome*+GO 105
5.4 Gene coverage retrieved by GSAn using DO, *Reactome*, GO or DO+*Reactome*+GO . 106
5.5 Example of annotation performed by GSAn when using DO and GO. 109
5.6 Box-plots of DO to GO mappings that were acquired by the *MAP_DO_GO*_1 and *MAP_DO_GO*_2 methods 111
5.7 Box-plots of mappings between *Reactome* and GO terms used before and after applying the GSAn method 112

6.1 Illustration of formal concept analysis representation 118

A.1 Percentage of covered genes for each gene set in the [C-260] dataset. 149
A.2 Percentage of covered genes for each gene set in the [B-346] dataset. 150

List of Tables

4.1 The nine semantic similarity measures investigated in this study. 67
4.2 List of organisms available in GSAn 85
4.3 Input parameters to be used in GSAn for the analysis 86
4.4 Comparison of gene set functional analysis tools. 89
4.5 Number of gene sets for which each tool provides an annotation. 90

5.1 Description of the structure of the three knowledge resources included in GSAn. 101
5.2 Mapping of DO and GO terms 107
5.3 Mapping of *Reactome* and GO terms 108

B.1 *Top ten* DO terms annotating the highest number of genes. 151
B.2 Information of the child terms of the *kidney cancer* DO term. 151
B.3 *Top ten Reactome* terms annotating the highest number of genes. 152

List of Abbreviations

AIC	**A**ggregate **I**nformation **C**ontent
ALHM	**A**verage **L**inkage **H**ierarchical **M**ethod
ASW	**A**verage **S**ilhouette **W**idth
BEL	**B**iological **E**xpression **L**anguage
BioNER	**Bio**medical **N**amed **E**ntity **R**ecognition
BioPAX	**B**iological **PA**thway e**X**change
BP	**B**iological **P**rocess
CC	**C**ellular **C**omponent
CCC	**C**ophenetic **C**orrelation **C**oefficient
CLHM	**C**omplete **L**inkage **H**ierarchical **M**ethod
DAG	**D**irected **A**cyclic **G**raph
DF	**D**istance **F**unction
DNA	**D**eoxy**R**ibonucleic **A**cid
DO	**D**isease **O**ntology
FCA	**F**ormal **C**oncept **A**nalysis
FCS	**F**unctional **C**lass **S**coring
FCT	**F**ind **C**ombined **T**erms
FIS	**F**requent **I**tem **S**et
GAF	**G**ene **A**ssociation **F**ile
GO	**G**ene **O**ntology
GOA	**G**ene **O**ntology **A**nnotation
GPAD	**G**ene **P**roduct **A**ssociation **D**ata
GPI	**G**ene **P**roducts **I**nformation
GSAn	**G**ene **S**et **An**notation
GSEA	**G**ene **S**et **E**nrichment **A**nalysis
HCL	**H**ue-**C**hroma-**L**uminance
HGNC	**H**uman genome organisation **G**ene **N**omenclature **C**ommite
HPO	**H**uman **P**henotype **O**ntology
IC	**I**nformation **C**ontent
ICD–10	**I**nternational **C**lassification of **D**iseases–10th version
IEA	**I**nferred from **E**lectronic **A**nnotation
IV	**I**nformation **V**isualisation
KEGG	**K**yoto **E**ncyclopedia of **G**enes and **G**enomes
LC	**L**eacock & **C**hodorow
LCA	**L**owest **C**ommon **A**ncestor
MDS	**M**ulti**D**imensional **S**caling
MEA	**M**odular **E**nrichment **A**nalysis
MeSH	**Me**dical **S**ubject **H**eadings
MF	**M**olecular **F**unction
MGD	**M**ouse **G**enome **D**atabase
MICA	**M**ost **I**nformative **C**ommon **A**ncestor
MOTVIS	**MO**dular **T**erm **VIS**ualization
MSigDB	**M**olecular **S**ignatures **D**ata**B**ase

MSRT	**M**ost **S**pecific **R**epresentative **T**erms
NCBI	**N**ational **C**enter for **B**iotechnology **I**nformation
NCIt	**N**ational **C**ancer **I**nstitute **t**hesaurus
ND	**N**o biological **D**ata available
NLP	**N**atural **L**anguage **P**rocessing
OBO	**O**pen **B**iomedical **O**ntologies
ORA	**O**ver-**R**epresentation **A**nalysis
ORDO	**O**rphanet **R**are **D**isease **O**ntology
OWL	**W**eb **O**ntology **L**anguage
PCA	**P**rincipal **C**omponents **A**nalysis
PS	**P**ekar & **S**taab
PT	**P**athway **T**opology based approaches
RDF	**R**esource **D**escription **F**ramework
RNA	**R**iboNucleic **A**cid
SA	**S**ynthetic **A**lgorithm
SCP	**S**et **C**over **P**roblem
SEA	**S**ingular **E**nrichment **A**nalysis
SGD	**S**accharomyces **G**enome **D**atabase
SKOS	**S**imple **K**nowledge **O**rganization **S**ystem
SLHM	**S**ingle **L**inkage **H**ierarchical **M**ethod
SMBL	**S**ystem **B**iology **M**arkup **L**anguage
SML	**S**emantic **M**easures **L**ibrary
SNOMED–CT	**S**ystematized **N**omenclature **O**f **M**edicine–**C**linical **T**erms
SO	**S**equence **O**ntology
TAIR	**T**he **A**rabidopsis **I**nformation **R**esource
TRANSFAC	**TRANS**cription **FAC**tor database
UMLS	**U**nified **M**edical **L**anguage **S**ystem

Dedicado a mi madre por ser una gran luchadora en todas las adversidades que le ha dado la vida y a la memoria de mi padre (1961-2019), que fue un trabajador, padre y una maravillosa persona…

Research contributions

This thesis has resulted in different research contributions: (i) journal papers, (ii) an article in the proceedings of an international conference, and (i) presentations during conferences.

Journal papers

1. GSAn: a Web server as an alternative to enrichment analysis for annotating gene sets.
   ***A. Ayllon-Benitez**, R. Bourqui, P. Thébault*[◕], *F. Mougin*[◕]

   NAR Genomics and Bioinformatics, *under review*

2. A new method for evaluating the impacts of semantic similarity measures on the annotation of gene sets.
   ***A. Ayllon-Benitez**, F. Mougin*[◕], *J. Allali, R. Thiébaut, P. Thébault*[◕]

   PLoS ONE 13(11): e0208037. DOI: 10.1371/journal.pone.0208037

International conference with published acts

1. Deciphering gene sets annotations with ontology-based visualization.
   20 min. talk
   ***A. Ayllon-Benitez**, P. Thébault, JT. Fernández-Breis, M. Quesada-Martinez, F. Mougin, R. Bourqui.*

   In 2017 IEEE 21st International Conferece on *Information Visualization* (IV), 11 – 14, July, London (UK). DOI: 10.1109/iV.2017.18

International conferences without published acts

Talks

1. rGSAn: An R package dedicated to the gene set analysis using semantic similarity measures.
   5 min. talk.
   ***A. Ayllon-Benitez**, F. Mougin, P. Thébault.*

   In 2019 useR! Conference, 9 – 12, July, Toulouse (France)

2. Evaluating the impacts of semantic similarity measures on the annotation of gene sets.
   20 min. talk.
   ***A. Ayllon-Benitez**, F. Mougin, P. Thébault.*

   In 2019 JOBIM Conference, 2 – 5, July, Nantes (France)

Posters

1. GSAn: An alternative to statistical analysis of gene sets.
   ***A. Ayllon-Benitez**, P. Thébault, F. Mougin.*

   In 2019 JOBIM Conference, 2 – 5, July, Nantes (France)

2. Deciphering functional annotation of multiple gene sets with MOTVIS.
   ***A. Ayllon-Benitez**, F. Mougin, M. Quesada-Martínez, J.T. Fernández-Breis, R. Bourqui, P. Thébault.*

   In 2018 JOBIM Conference, 3 – 6, July, Marseille (France)

[◕] equal contribution

Résumé substantiel

1 Introduction

L'émergence de nouvelles technologies de séquençage a fortement influencé notre compréhension des relations entre génotype (collection de gènes) et phénotype (caractéristiques observables codées par ces gènes). Les avancées dans l'analyse de l'expression différentielle de gènes ont suscité un vif intérêt pour l'étude des ensembles de gènes présentant une similarité d'expression dans une même condition expérimentale. Les approches classiques pour interpréter l'information biologique reposent sur l'utilisation de méthodes statistiques. Cependant, ces méthodes se focalisent sur les gènes les plus connus tout en générant de la redondance d'information qui peut être éliminée en prenant en compte la structure des ressources de connaissances qui fournissent l'annotation. Au cours de cette thèse, nous avons exploré différentes méthodes permettant l'annotation d'ensembles de gènes avec l'objectif d'améliorer la compréhension de leur contexte biologique.

2 Resources de connaissances et annotation fonctionnelle

Les informations relatives à un gène ou produit de gène se présentent sous la forme d'annotations (associations entre les gènes et ses phénotypes, obtenues par des méthodes expérimentales ou computationnelles). Un gène peut être associé à différentes informations biologiques (désignées par le terme d'*annotation fonctionnelle*) telles que sa localisation dans un génome donné, la description de l'activité de ses produits (tels que les protéines) ou ses fonctions spécifiques dans une cellule.

Tous ces informations sont stockées dans différents types de ressources de connaissances telles que les bases de connaissances, thésauri ou ontologies. Parmi le grand nombre de ressources de connaissances disponibles en bioinformatique, la *Gene Ontology* (GO) est celle qui est la plus largement utilisée [Con04a; Con18]. GO est une ressource décrivant en particulier les processus et fonctions des gènes et produits de gènes. Les concepts de GO (dénommés "termes" dans GO) sont organisés sous la forme d'un graphe orienté acyclique qui compte plus de 44000 termes connectés par différents types de relations (*e.g.*, *is_a*, *part_of*, *regulates*). Un des intérêts de GO est l'existence de bases de données d'annotation (la *GO Annotation* ou GOA), qui contiennent les associations entre les gènes de plusieurs organismes et les termes GO qui les annotent. Ces associations, ou **annotations**, sont générées de diverses manières (*e.g.*, par des méthodes expérimentales, automatiques). À chaque annotation est associée un code d'évidence qui décrit la manière précise dont elle a été obtenue.

De nombreuses ressources de connaissances existent, notamment sur le portail biomédical du *National Center for Biotechnology Information* (NCBI) [OLe+15] qui fournit des informations et des services facilement accessibles. En se concentrant sur l'*annotation fonctionnelle*,

un gène peut être impliqué dans d'autres situations que des processus ou des fonctions biologiques. Par exemple, les gènes peuvent également être la cible de médicaments, être exprimés dans diverses voies métaboliques, ou même être des précurseurs de maladies. Dans le cadre de cette thèse, en plus d'utiliser GO et GOA pour l'annotation d'ensembles de gènes, nous nous sommes intéressés aux rôles joués par les gènes dans les voies métaboliques et les maladies.

Malgré l'utilité de ces ressources pour déterminer les rôles biologiques d'un gène, les informations ne suffisent pas à comprendre une condition expérimentale donnée après avoir comparé deux populations (*e.g.*, une population saine et une autre atteinte d'une maladie). La compréhension d'une maladie donnée, ainsi que l'étude de l'impact d'un médicament ou d'un vaccin impliquent l'interaction de plusieurs gènes à un moment précis. Il est plus intéressant de savoir quels processus biologiques sont produits par un ensemble de gènes que de comprendre les fonctions d'un seul gène. À l'heure actuelle, les études portent sur des ensembles de gènes qui sont souvent sur-exprimés au même moment puisqu'ils coexistent dans la même voie métabolique [Sub+05]. L'utilisation d'approches par ensembles de gènes est donc devenue de plus en plus populaire. De nombreux travaux ont été consacrés à la création, à la gestion et à la conservation d'ensembles de gènes [Ant11; Bak+11; Cul+11; Li+13; CB14; Lib14].

Ce nouveau champ de recherche est devenu incontournable au cours des deux dernières décennies et repose sur la création d'ensembles de gènes à partir d'une comparaison de résultats expérimentaux dans diverses conditions. Par exemple, Chaussabel et Baldwin [CB14] ont mis en œuvre des conditions expérimentales liées à diverses maladies pour déchiffrer les gènes pouvant être impliqués dans la réponse immunitaire innée ou acquise (spécifique). Cependant, l'interprétation d'un ensemble de gènes n'est pas une tâche aisée. Étant donné que chaque gène dispose de ses propres annotations, la grande quantité d'information à traiter pour un ensemble de gènes rend sa compréhension difficile. À titre d'exemple, si on considère que les gènes humains sont annotés en moyenne par 10 termes GO, un ensemble de 100 gènes peut produire plusieurs centaines de termes dont le sens peut se recouper, générant dans ce cas de la redondance. De plus, lorsque l'on étudie plusieurs ensembles hétérogènes de gènes, le nombre de termes peut même atteindre plusieurs milliers. Ainsi, l'expertise manuelle pour décrypter clairement les principales fonctions qui peuvent être liées au(x) ensemble(s) de gènes étudié(s) est chronophage et devient ingérable lorsque le nombre d'ensembles de gènes augmente. Pour ces raisons, des méthodes automatiques ont été proposées pour faciliter l'analyse d'ensembles de gènes. Nous nous sommes intéressés à quatre domaines d'étude qui ont pour objectif d'annoter un ensemble de gènes ou d'améliorer son interprétation.

1. **Annotation par des techniques d'enrichissement classiques.** Diminuer le nombre de termes d'annotation tout en conservant les plus informatifs est un défi majeur pour comprendre les implications biologiques d'un ensemble de gènes [BPG16]. L'une des approches classiques pour interpréter l'information biologique liée à un ensemble de gènes est l'enrichissement. Le principe de ces méthodes statistiques est de comparer, sur la base de ses annotations, un ensemble de gènes à une référence (*e.g.*, le génome complet ou un ensemble de gènes généré de manière aléatoire dont la taille est comparable à l'ensemble de gènes étudié) pour trouver les termes d'annotation sur-représentés au sein de l'ensemble en question (*i.e.*, utilisés plus que la "normale" par les gènes de

l'ensemble). Un grand nombre d'outils implémentant des approches d'enrichissement ont été développés. Cela permet de fournir une annotation à un ensemble de gènes (voir détails et classification dans les revues de Huang et al. [HSL09] et Khatri et al. [KSB12]). Cependant, ces méthodes d'enrichissement fournissent une liste de termes sur-représentés sans considérer la pertinence de cette information ni la spécificité des termes d'annotation obtenus. De plus, les annotations proposées par ces méthodes présentent une certaine redondance par la présence de termes reliés hiérarchiquement dans une ressource de connaissances. Enfin, ces méthodes tendent à se concentrer sur les gènes les plus étudiés et fournissent des résultats d'annotation couvrant un nombre limité des gènes de l'ensemble [BLG15; HTK18; Tom+18].

2. **Avantages de l'annotation ontologique.** Le fait que les annotations associées aux gènes soient issues d'une ontologie est un avantage. En particulier, chaque concept d'une ontologie possède des caractéristiques qui peuvent être exploitées pour déterminer sa similarité avec d'autres concepts. De cette manière, les caractéristiques partagées par deux termes GO peuvent être prises en compte pour quantifier leur similarité, et donc leur redondance ou leur complémentarité. Il existe de nombreuses mesures de similarité sémantique[1] [Pes+09; Guz+12; MCM17], rendant le choix d'une mesure par rapport à une autre délicat.

3. **Exploration des résultats d'annotation avec la visualisation.** La visualisation est très utile pour explorer des connaissances. Le nombre de techniques de visualisation utilisées dans le domaine biologique a considérablement augmenté au cours des 15 dernières années [Ker+17]. Cependant, la visualisation est encore à un stade précoce où il reste des défis importants à relever liés en particulier au volume des données, à leur type et à leur représentation [ODo+10; Mou+18]. Une combinaison de différentes visualisations est une solution indiquée pour représenter des résultats divers avec plusieurs niveaux d'information. Cependant, cet usage reste rare en biologie. De plus, les méthodes actuelles utilisées pour la visualisation de termes d'annotation d'un ensemble (réduit) de gènes ne sont pas adaptées pour traiter plusieurs dizaines d'ensembles de gènes.

4. **Exploiter plusieurs ressources de connaissances pour améliorer la compréhension biologique.** L'ensemble des connaissances du domaine biologique ne se trouvent pas dans une unique ressource. De plus, certaines ressources de connaissances peuvent décrire des notions similaires (voire identiques), générant de la redondance et potentiellement des contradictions [KPL03]. Il est donc essentiel d'intégrer des ressources dont les connaissances représentées se recoupent, mais qui sont aussi complémentaires, afin d'unifier l'information et d'être en mesure de fournir une vue d'ensemble des connaissances biologiques. Selon Keet [Kee04], le concept d'*intégration* signifie tout ce qui concerne la fusion, l'utilisation, le mapping[2], l'extension, l'approximation, les vues unifiées entre deux éléments (*e.g.*, deux ensembles de données, deux ressources de connaissances, deux visualisations). Notons que Manzoni et al. [Man+16] mentionnent d'importants défis liés à l'intégration, notamment l'existence de multiples ressources

[1] La *similarité sémantique* consiste à trouver en quoi deux concepts sont proches du point de vue de leur sens en utilisant des caractéristiques propres aux ressources de connaissances dans lesquels ils sont décrits.

[2] Le terme *mapping* correspond à une correspondance entre deux éléments. Si une correspondance existe, on dit que ces deux éléments sont *mappés*.

de connaissances pour décrire la même chose dans différentes bases de données biologiques, l'origine des données ou la capacité de stockage due à l'intégration.

Dans cette thèse, nous proposons des solutions informatiques visant à résoudre la plupart des défis présentés ci-dessus.

3 Visualisation de l'annotation des ensembles de gènes

En raison de la taille et de la complexité des données d'annotation, il est nécessaire de recourir à des techniques de visualisation adaptées à l'interprétation de ces informations. Cependant, le choix de la métaphore visuelle adéquate est une tâche difficile.

Le premier chapitre présente les solutions visuelles que nous avons proposées pour faciliter l'interprétation des résultats d'annotation d'un ou plusieurs ensembles de gènes. Pour cela, nous avons proposé des solutions pour visualiser des résultats d'annotation d'un ensemble de gènes comparativement à plusieurs ensembles. De plus, nous avons également étudié l'impact d'utiliser ou non la structure des ressources de connaissances fournissant l'annotation pour améliorer l'interprétation des résultats. Nous avons ainsi développé trois prototypes de visualisation.

- Les deux premiers prototypes visaient à représenter l'annotation fonctionnelle d'un seul ensemble de gènes sans tenir compte de la structure de la ressource de connaissances d'où sont issus les termes d'annotation. Ces métaphores visuelles sont le résultat de développements réalisés par des étudiants (de master 1 et première année d'ingénieur) que j'ai encadrés. Ce travail exploratoire a eu le double avantage de me permettre d'étudier des métaphores visuelles pour explorer l'annotation (obtenue par des méthodes d'enrichissement) d'un ensemble de gènes et d'expérimenter l'encadrement d'étudiants.
- Le troisième outil de visualisation est le résultat d'une collaboration internationale avec l'Université de Murcie (Espagne). Il visait à réconcilier les annotations issues de différentes ressources de connaissances utilisées pour annnoter plusieurs ensembles de gènes en utilisant des similitudes lexicales entre les termes d'annotation. Cette visualisation est une combinaison de deux métaphores visuelles prenant en considération la structure hiérarchique de GO pour l'affichage et l'exploration des résultats.

3.1 Visualisation de l'annotation d'un ensemble de gènes

Pour cette première exploration, un processus d'analyse a été conçu. Celui-ci combine des méthodes statistiques d'enrichissement, des mesures de similarité sémantique, de clustering[3] et la sélection du terme parent le plus informatif (connu en anglais comme *MICA* pour *Most Informative Common Ancestor*) pour chaque cluster de termes. Les résultats ont pu être affichés grâce à deux prototypes de visualisation : (i) un prototype de visualisation montrant les différents clusters et les termes GO s'y trouvant, et (ii) un prototype qui représente les relations entre les gènes, les termes GO et les termes MICA. Pour pouvoir explorer les résultats

[3] Le terme *clustering* est utilisé pour décrire une méthode qui regroupe des éléments ayant des caractéristiques similaires. En appliquant cette méthode sur les termes GO, les groupes de termes seront définis comme des *clusters* de termes GO

selon plusieurs niveaux de profondeur, nous avons ajouté des fonctionnalités d'interaction aux prototypes. Pour illustrer l'intérêt de la visualisation que nous avons proposée, nous avons effectué une analyse sur un ensemble de 27 gènes annotés avec le terme "interféron" par Chaussabel et Baldwin [CB14].

Nous avons ainsi observé que la visualisation des termes MICA permet de résumer l'information.. De plus, la vue d'ensemble du deuxième prototype de visualisation donne une synthèse claire des résultats à l'aide des termes MICA, tandis que la première exige que les utilisateurs explorent en détails les clusters de termes GO similaires pour extraire l'information pertinente. Nous avons constaté par ailleurs que les méthodes d'enrichissement d'un ensemble de gènes génèrent une annotation avec un degré élevé de redondance parmi les termes GO, comme décrit précédemment.

3.2 Visualisation de l'annotation de plusieurs ensembles de gènes

Un deuxième processus a été conçu pour annoter plusieurs ensembles de gènes. Celui-ci se base sur les résultats de méthodes d'enrichissement et exploite plusieurs ressources de connaissances pour obtenir des termes d'annotation. Nous avons utilisé l'outil `OntoEnrich` [Que+15a] pour aligner des termes issus d'autres ressources de connaissances aux termes GO grâce à des mesures de similarité lexicale[4]. Ensuite, la structure de GO a été simplifiée pour faciliter l'exploration des résultats d'annotation. Ces résultats peuvent être explorés via un prototype de visualisation que nous avons développé, `MOTVIS`, qui combine deux vues inter-connectées : une arborescence qui fournit une vue d'ensemble mais aussi des informations détaillées sur les ensembles de gènes, et une visualisation qui permet de se concentrer sur les termes d'intérêt de l'annotation. `MOTVIS` offre aussi des fonctionnalités d'interaction permettant une exploration des résultats en profondeur.

Pour illustrer l'utilité de notre processus, nous avons réalisé une analyse en utilisant les données fournies par Chaussabel et Baldwin [CB14] sous la forme d'un répertoire de 260 ensembles de gènes concernant la réponse immunitaire au sein de la population humaine. Les conclusions majeures de ces analyses sont que : (i) l'analyse lexicale réalisée par `OntoEnrich` a été utile pour éliminer les redondances entre les termes décrivant la même notion dans des ressources de connaissances différentes, et (ii) l'utilisation de la structure de GO au sein de la visualisation permet de réduire la complexité visuelle et facilite ainsi l'interprétation des résultats.

4 Annotation d'un ensemble de gènes à l'aide de mesures de similarité sémantique

La plupart des outils classiques pour annoter des ensembles de gènes reposent sur des méthodes d'enrichissement statistique qui comportent généralement deux étapes : une étape *a priori* qui vise à synthétiser l'annotation en sélectionnant les termes sur-représentés, et une étape *a posteriori* qui élimine les informations potentiellement redondantes en utilisant la structure des ressources de connaissances décrivant les termes d'annotation. Cependant,

[4] La *similarité lexicale* consiste à estimer la similarité entre deux concepts dont les noms se ressemblent lexicalement.

même en proposant une étape *a posteriori*, les méthodes statistiques d'enrichissement mettent inévitablement en évidence les gènes les plus étudiés au détriment des gènes mal annotés lors de l'analyse [BLG15; HTK18; Tom+18], ce qui entraîne une perte d'information.

Dans ce travail, nous avons abordé ces limites en proposant une méthode originale qui annote les ensembles de gènes en utilisant des mesures de similarité sémantique. À notre connaissance, aucun travail n'a été proposé pour évaluer l'impact de l'utilisation d'une mesure de similarité sémantique donnée par rapport à une autre mesure quand elles sont utilisées pour annoter un ensemble de gènes. Une annotation pertinente d'un ensemble de gènes doit répondre à des caractéristiques spécifiques pour fournir des informations utiles aux experts du domaine.

4.1 Impact de la similarité sémantique dans l'annotation d'un ensemble de gènes

Nous avons évalué l'impact de chaque mesure en considérant les critères suivants pour définir une "bonne" annotation d'un ensemble de gènes [BPG16]:

- Le nombre de termes d'annotation doit être considérablement réduit tout en garantissant qu'ils représentent correctement l'ensemble de gènes (critère désigné sous le nom de **synthèse**).
- Le nombre de gènes décrits par les termes sélectionnés (critère désigné sous le terme de **couverture**) doit être maximisé.

Le défi consiste alors à trouver le meilleur compromis possible entre les deux critères tout en s'efforçant de maintenir un niveau de détails suffisant apporté par les termes choisis. Nous avons mis en place une méthode visant à étudier l'impact d'utiliser différentes mesures de similarité sémantique sur l'interprétation d'un ensemble de gènes.

À partir de la totalité des termes GO extraits pour chaque gène de l'ensemble étudié, un premier filtre a été appliqué pour supprimer les annotations n'apportant pas d'information pertinente pour le gène (*i.e.*, les annotations redondantes ou incomplètes).

Afin d'étudier l'impact d'utiliser différentes mesures de similarité sémantique, nous avons sélectionné neuf mesures parmi les trois classes suivantes, définies par Pesquita et al. [Pes+09] : mesures *basées sur les nœuds*, mesures *basées sur les arêtes* et mesures *hybrides*. Les mesures *basées sur les nœuds* calculent la similarité entre deux termes à partir des propriétés spécifiques aux termes, comme leur profondeur ou leur contenu d'information (*CI*). Les mesures *basées sur les arêtes* exploitent la distance qui sépare deux termes GO au sein du graphe GO. Les mesures *hybrides* utilisent, quant à elles, une combinaison des deux types précédents.

Ensuite, nous avons examiné la capacité de chaque mesure de similarité sémantique à obtenir les meilleures partitions des termes d'annotation en évaluant la pertinence des partitions du clustering et l'impact des différentes méthodes de clustering.

Pour identifier les termes les plus synthétiques de l'ensemble de gènes tout en évaluant la combinatoire des solutions, nous avons développé un algorithme de parcours de la structure de GO afin de récupérer tout d'abord un ou plusieurs termes (ce nombre étant dépendant de

la taille du cluster) représentatifs pour chaque cluster de termes obtenu. Nous avons ensuite examiné la capacité de chaque mesure de similarité sémantique à : (i) réduire le nombre de termes d'annotation tout en sélectionnant les termes les plus représentatifs de l'ensemble de gènes, et (ii) fournir une annotation synthétique incluant le plus grand nombre de gènes possible. La combinaison de ces deux critères, essentiellement quantitatifs, a permis d'estimer la capacité de chaque mesure de similarité sémantique à produire une annotation plus pertinente et synthétique pour un ensemble de gènes donné.

Nous avons utilisé deux jeux de données de l'organisme "homo sapiens" qui contiennent respectivement 260 et 346 ensembles de gènes liés à la réponse immunitaire pour évaluer notre méthode. Les différentes évaluations sur les partitions de clustering ont montré de bons résultats avec les mesures *basées sur les nœuds* par rapport à celles *basées sur les arêtes* et de mauvais résultats pour les mesures *hybrides*. Pour étudier l'impact de chaque mesure de similarité sémantique sur l'obtention d'une annotation synthétique, nous avons analysé la quantité de termes retenus et le nombre de gènes couverts en observant la pertinence en terme d'information biologique (mesure basée sur le *CI*). En comparant ces résultats avec l'outil d'enrichissement `DAVID` [Den+03], nous avons montré que notre approche donnait de meilleurs résultats pour la majorité des mesures de similarité sémantique, les mesures de similarité sémantique *basées sur les nœuds* étant les meilleures. Les mesures de similarité sémantique permettent ainsi de trouver un bon équilibre pour garantir une meilleure couverture du nombre de gènes avec un nombre minimum de termes (tout en gardant une information pertinente et synthétique).

4.2 Extension du cadre d'analyse

À ce stade, nous avons développé une méthode fournissant une annotation synthétique pour un ensemble de gènes donné. Dans cette section, nous décrivons des étapes supplémentaires enrichissant la méthode précédente afin d'améliorer la qualité de cette annotation. Parmi ces étapes, une algorithme basé sur **le problème de couverture par ensembles** [VLZ16] a été conçu pour sélectionner les termes les plus synthétiques sans affecter la couverture de gènes fournie par les termes représentatifs.

À partir de ces améliorations de la méthode, nous avons proposé un serveur web, `GSAn` (`https://gsan.labri.fr`), qui offre un approche alternative aux méthodes d'enrichissement pour l'annotation d'un ensemble de gènes. `GSAn` exploite donc des mesures de similarité sémantique afin de réduire l'annotation *a priori*. Par ailleurs, `GSAn` offre une visualisation originale et interactive facilitant l'interprétation des résultats par les experts qui peuvent choisir le niveau d'information biologique qui leur semble pertinent. `GSAn` permet aux utilisateurs d'annoter une liste de symboles de gènes ou de protéines et fournit un ensemble de composantes visuelles favorisant la compréhension des résultats d'annotation : (i) trois diagrammes circulaires présentant des informations générales sur l'ensemble de gènes (couverture par GOA et `GSAn` et similarité des termes GO au sein de l'ensemble), (ii) un diagramme en barres montrant les informations concernant les termes synthétiques, (iii) un tableau représentant les informations des termes représentatifs, et (iv) la combinaison de visualisations arborescentes implémentée dans `MOTVIS` pour présenter conjointement les termes représentatifs et les gènes

de l'ensemble étudié.

Pour illustrer l'intérêt de GSAn, nous avons aussi étudié les jeux de données de 260 et 346 ensembles de gènes issus d'une approche de transcriptomique étudiant la réponse immunitaire. Deux analyses ont été réalisées avec GSAn : (i) une comparaison des résultats d'annotation de GSAn par rapport à des outils d'enrichissement, et (ii) l'étude d'un ensemble de gènes qui a été annoté par des experts comme *régulation de la présentation d'antigènes* et *réponse immunitaire*. GSAn a montré d'excellents résultats en termes de maximisation de la couverture des gènes tout en minimisant le nombre de termes. GSAn a fourni une annotation plus spécifique que les résultats donnés par les experts pour l'ensemble de gènes étudié. De plus, GSAn présente l'avantage de fournir des capacités de visualisation interactive pour analyser les annotations de l'ensemble de gènes. La visualisation MOTVIS offre, quant à elle, une diversité d'interactions permettant de parcourir les termes et les gènes qu'ils annotent en fonction du niveau d'information biologique qui peut intéresser les utilisateurs.

5 Intégration de ressources de connaissances supplémentaires au sein de GSAn

La dernière étape de cette thèse s'est intéressée à la question de la faisabilité d'intégrer d'autres ressources de connaissances dans GSAn. Nous avons ainsi intégré deux ressources, l'une décrivant les maladies humaines et l'autre les voies métaboliques, afin de fournir des informations supplémentaires aux utilisateurs. Notre objectif était d'améliorer la couverture des gènes annotés sans pour autant augmenter significativement le nombre de termes synthétiques fournis par GSAn.

Concernant les maladies, nous avons considéré que *Disease Ontology* (DO) était un candidat idéal pour les raisons suivantes : (i) elle couvre un large éventail de maladies, (ii) il est possible de récupérer des associations entre les termes DO et les gènes, et (iii) cette ressource de connaissances contient des références croisées vers d'autres ressources telles que SNOMED-CT, MeSH, HPO et ORDO. DO contient 9384 termes uniques décrivant des maladies humaines.

Concernant les voies métaboliques, nous nous sommes concentrés sur *Reactome* car cette ressource de connaissances a l'avantage de fournir : (i) un accès facile via une interface web, (ii) de nombreux formats de données, et (iii) un alignement avec GO. *Reactome* contient plus de 2300 voies uniques impliquant 16 espèces. Parmi les données de *Reactome*, nous nous sommes focalisés sur les voies métaboliques pour analyser l'apport de ce type d'information au sein de GSAn.

GSAn manipule des formats spécifiques pour GO et GOA. Par conséquent, il a été nécessaire de convertir *Reactome* et DO afin qu'elles puissent être incluses dans GSAn. Les associations entre les gènes et les termes de maladie ne sont pas officiellement fournies par DO. Cependant, elles peuvent être extraites de bases de données externes [Piñ+15; Ple+15] en utilisant notamment un algorithme implémenté dans INTEGRO [CGV18]. Cet outil exploite les références croisées des ressources de connaissances externes fournies par DO pour récupérer les gènes.

Ensuite, nous avons évalué l'impact de l'ajout de ces deux ressources de connaissances dans GSAn lors de l'analyse d'ensembles de gènes. Dans cette première analyse, aucune intégration n'a été effectuée. Ainsi, nous avons appliqué GSAn en nous basant sur chaque ressource de connaissances indépendamment et nous avons ensuite fusionné les résultats générés. Nous avons démontré que l'intégration d'autres ressources de connaissances dans GSAn a amélioré la couverture des gènes annotés. Néanmoins, l'utilisation séparée des différentes ressources de connaissances a généré une augmentation du nombre de termes d'annotation, ce qui pourrait rendre les résultats plus difficiles à interpréter.

Pour résoudre ce problème, deux stratégies d'intégration ont été étudiées dans un deuxième temps : (i) un mapping *a priori*, et (ii) un mapping *a posteriori* du calcul de l'annotation par GSAn. À notre connaissance, il n'existe pas de mapping entre GO et DO. Pour résoudre ce problème, trois méthodes ont été mises au point, puis comparées. Deux méthodes utilisent le même principe qui consiste à considérer les gènes associés à un terme DO comme un ensemble de gènes. Ensuite, ces ensembles ont été analysés par une méthode d'enrichissement (méthode 1) ou par GSAn (méthode 2) afin de récupérer l'annotation GO pour ces ensembles. Ainsi, le mapping implique qu'un terme DO soit mappé avec un ou plusieurs termes GO. La dernière méthode combine l'annotation de gènes fournie par GO et DO. Pour ce faire, nous avons implémenté une étape préliminaire pour déduire des *règles d'association*, basée sur l'algorithme des ensembles d'éléments fréquents (ou *frequent item set* en anglais) [Nau+13]. Cela correspond à un ensemble d'éléments qui apparaissent fréquemment lorsqu'un modèle de données est considéré. Ces mappings ont ainsi été établis en exploitant les instances, à savoir les gènes associés aux termes des ressources de connaissances.

Nous avons finalement étudié les jeux de données de 260 et 346 ensembles de gènes mentionnés précédemment. Comme attendu, l'intégration *a priori* s'est révélée plus pertinente car elle a permis de trouver davantage de mappings entre une ressource de connaissances et GO pour un ensemble de gènes donné. De plus, pour établir les mappings entre les termes de DO et GO, la méthode exploitant *GSAn* a fourni une meilleure couverture de mappings entre ressources que les autres méthodes.

6 Conclusions et perspectives de recherche

Nous avons décrit une synthèse des quatre objectifs de cette thèse pour annoter ou améliorer l'annotation d'un ensemble de gènes donné. Ces objectifs étaient les suivants :

1. Annoter des ensembles de gènes par une analyse d'enrichissement présentant la sur-représentation des termes d'annotation.

2. Améliorer l'annotation sur-représentée avec des solutions exploitant les caractéristiques de termes d'annotation issue d'une ontologie pour les grouper dans des clusters en fonction de leur similarité sémantique.

3. Représenter les résultats d'annotation d'un ou plusieurs ensembles de gènes au sein de plusieurs métaphores visuelles afin de mieux comprendre un contexte biologique donné.

4. Mettre en œuvre des techniques d'intégration pour proposer des annotations issues de différentes ressources de connaissances afin d'enrichir l'information biologique caractérisée par un ensemble de gènes.

Au cours de nos travaux, `GSAn` a été développé comme alternative aux outils basés sur l'enrichissement afin d'annoter un ensemble de gènes et proposant une visualisation interactive permettant d'explorer en détail les résultats d'annotation d'un ensemble de gènes.

Même si `GSAn` est disponible sur Internet et produit de bons résultats d'annotation, il reste des améliorations à apporter pour fournir des annotations encore plus riches et précises. Parmi les perspectives possibles, on peut citer l'intérêt : (i) d'inclure d'autres ressources de connaissances, décrivant par exemple les relations entre gènes dans un *réseau génique de régulation*, et (ii) d'améliorer l'étape d'intégration de nouvelles ressources dans `GSAn`.

Chapter 1

Introduction

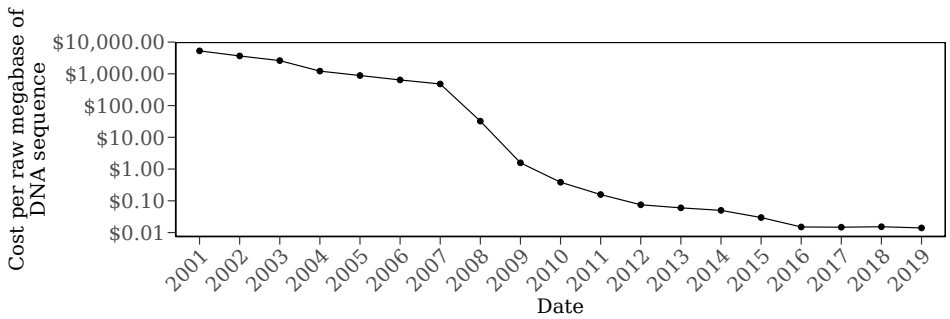
1.1 The role of bioinformatics into biological and clinical researches

Bioinformatics arose from the need to use computational methods for solving problems in biology. Originally based on interactions between chemists, physicists, statisticians, mathematicians, and biologists to create applications for biology, bioinformatics is nowadays a vast domain following its own lines of research. Even if the bioinformatics domain is relatively recent, its origins date back to more than 50 years ago (see a quick resume along the brief history in Gauthier et al. [Gau+18]). Key events where bioinformatics had an important role include the Human Genome Project [Lan+01; Con04b], the Human Microbiome Project [Tur+07] and, recently, the 1,000 Genomes Project [Con+15]. The emergence of the Internet in the nineties allowed developing many bioinformatics resources accessible to everyone throughout the world. These resources allow to explore, manipulate and generate biological information for understanding diverse biological questions.

Bioinformatics covers a large spectrum of research fields such as genomics, proteomics, metabolomics, phylogenetics, transcriptomics, high-throughput image analysis, and drug design. Today, with the development of high-throughput technologies and the massive generation of data to be treated, bioinformatics faces new challenges, including the management of big data, the reproducibility of results and the integration of different types of biological data [Gau+18].

Advances in the bioinformatics field, particularly within the omics field, paved the way for personalized medicine. Recently, terms like *clinical bioinformatics*, *translational bioinformatics* or *genomic medicine* have emerged expressing the need for combining bioinformatics and clinical sciences [But08; Sar+11; WL11; Bel+12]. These clinical fields may expect that bioinformatics methodologies originally focused on biological discoveries can now be applied to improve human health. Because performances of high-throughput sequencing machines increase over the years, the cost per megabase of DNA (*DeoxyRibonucleic Acid*) sequences (figure 1.1A) and per genome (figure 1.1B) has drastically decreased. This decrease makes information related to gene expression profiling (measurement of the activity of genes in a given condition) more accessible for being used in clinical studies. The gene expression profiling may help to better understand how genes are involved in particular biological or biomedical conditions (*i.e.*, a given biological process in a particular disease).

(A)



(B)

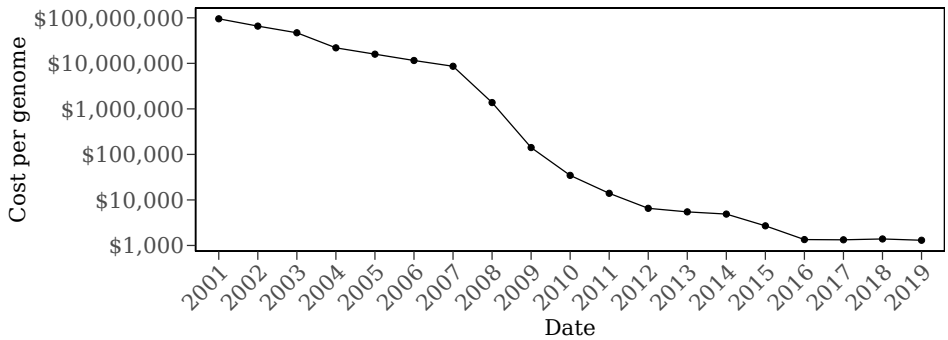


FIGURE 1.1: Decreasing tendency in (A) megabase DNA sequences' cost and (B) genome cost from September 2001 to February 2019. The data are available on the National Human Genome Research Institute's web site (https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data).

1.2 Motivation to improve the information ciphered by gene sets

The discovery of new medical drugs or vaccines implies to carry out large scale testing of new therapies. The evaluation of their impacts requires to compare experimental data from at least two populations (or one population over the time): one for the control cases (*e.g.*, healthy population, population before receiving the vaccine) and one for a population having a chronic disease (or the same population after receiving the vaccine). Over the past decade, the revolution in new sequencing technologies has strongly supported the production of omics data to improve our understanding of the relations between genotype (collection of genes) and phenotype (the observable features that are encoded by those genes) while comparing two populations. Making use of these datasets is crucial to identify the genes involved in a experimental condition by comparing the expression levels of genes in both populations. Nowadays, a common method consists in comparing the expression levels of genes in both populations acquired using next-generation sequencing techniques. The key idea of such studies is then to identify groups of impacted genes having common biological functions to better understand their roles in the phenotype.

Such studies consider groups of genes that are often related as they co-occur in the same pathway [Sub+05]. Several genes that are differentially expressed in a given pathway are more relevant than only one gene with a strongly different expression pattern. These groups,

hereinafter referred as **gene sets**, are formed based on experimental data by using clustering algorithms or statistical approaches [JTZ04]. Then, to decipher the biological functions or processes of these gene sets, it is necessary to access the information associated with genes stored in knowledge resources. At last, researchers are interested in understanding the different cell functions that may be impacted by a vaccine, a drug or a disease. Considering human data, users can obtain much information describing the gene functions according to different databases.

Consequently, the amount of information describing the biological function of genes in a particular set might be too large for being treated manually. For example, focusing on *Gene Ontology Annotation* (GOA) [Cam+04] (one of the most popular annotation database), on average, each human gene is described by the use of 10 biological terms, giving information with varying degrees of details. At the gene level, the attribution of terms to a gene mostly requires computational methods that make use of experimental and proved information for specific genes to transfer to genes they can be related to, from an evolutionary point of view. At the gene set level, automatic enrichment methods [HSL09] make use of statistical approaches from lists of terms (coming from each gene) to reduce the number of terms, by keeping those that are relevant for the investigated gene set. These methods, combined with visualizations for exploring results, are widely accepted and used by biologists and clinicians. Nevertheless, enrichment methods have drawbacks recently reported by several authors [BLG15; HTK18; Tom+18]. Among these limitations, it should be noted that these methods do not consider the inter-relations between terms of a given knowledge resource or across multiple resources. This may generate redundancies within results and loss of precise or synthetic information that may describe relevant information for a particular gene set.

During this thesis, we explored different study fields to improve the understanding of the biological context of gene sets. Our objective was to solve some drawbacks of current methods performing gene set annotation and to provide a synthetic annotation whose precision and relevance allow biologists or clinicians to better understand their gene sets. Hence, we used knowledge resources called ontologies, which organize their entities according to their semantic meanings, and in particular the *Gene Ontology* (GO) [Ash+00] that is one of the most famous and exhaustive biological ontologies.

1.3 Project outline

This thesis is structured according to **six chapters**. After this introductory chapter, the rest of this manuscript is organized as follows:

> In **chapter 2**, existing works associated with annotation at the gene level and at the gene set level are presented. At the gene level, we will introduce knowledge resources that describe information about genes, by emphasizing on the GO and GOA. At the gene set level, since the amount of information considerably increases, we will describe different solutions that provide relevant annotations as well as their drawbacks that we tried to address in the works presented in the following chapters.

> **Chapter 3** is dedicated to the visualization of gene set annotation results. In particular, we will study the impact of visualizing annotations provided by enrichment analysis for a single gene set and multiple gene sets. For a single gene set, we developed a pipeline based on enrichment analysis for exploring results using two different visual metaphors. For the analysis of multiple gene sets, we collaborated with a Spanish research group to annotate all gene sets using different knowledge resources. Then, we merged results for displaying them according to the structure of a unique knowledge resource thanks to a visualization prototype that we called `MOTVIS` (*MOdular Term VISualization*).

> **Chapter 4** is focused on the annotation of a gene set by using an alternative method to classical approaches based on enrichment analysis. First, we propose a new method to annotate a gene set based on semantic similarity between annotation terms, clustering methods and a new combinatorial method to provide a list of relevant terms to gene sets. Following this strategy, we evaluated the impact of using different semantic similarity measures by taking into consideration the two following features that correspond to relevant criteria for characterizing a *good* synthetic gene set annotation: (i) the number of annotation terms has to be drastically reduced while (ii) the representative terms maintain a sufficient level of details and (iii) the number of genes described by the selected terms should be as large as possible. Second, according to the results obtained in the first study, we developed `GSAn` (*Gene Set Annotation*), a novel gene set annotation web server that uses semantic similarity measures to synthesize *a priori* GO terms. Moreover, `GSAn` offers interactive visualization facilities dedicated to the multi-scale analysis of gene set annotations.

> To go further and extend `GSAn` to other knowledge resources, **chapter 5** presents an early stage of the integration of additional knowledge resources. To do that, we included two new resources to be combined with Gene Ontology describing pathways and diseases. In particular, we include these new knowledge resources to `GSAn` and search for mappings to find equivalences between these resources and GO.

> Finally, in **chapter 6**, we will conclude this manuscript and present some perspectives for future researches.

1.4 Datasets used in this thesis

To evaluate the efficiency of methods developed along this thesis, we used at least one the following two datasets:

[`C-260`] The `Chaussabel and Baldwin` dataset comprises 260 genes sets, that were computed according to their expression profile achieved from one to fifteen experimental protocols using a co-expressed network approach [CB14]. Each of these protocols corresponds to a study of a disease to identify genes that trigger an immune response. More precisely, the modules that gather genes with a similar behavior within the fifteen experimental protocols may be related to the immune response in general. In contrast, modules of genes that show similar behavior in a limited number of protocols may be expressed specifically in some diseases.

[B-346] The BTM (blood transcriptional modules) dataset comprises 346 gene sets and aims to characterize innate and adaptive immune responses in vaccination studies [Li+13]. The gene sets were built through large-scale data integration of publicly available transcriptomes of human blood using "interactome", "bibliome", pathway databases and specific biological contexts to deduce a set of transcriptional modules.

The [C-260] dataset has been used in works described in chapter 3 while [C-260] and [B-346] were used in chapter 4 and chapter 5. These datasets are focused on human genes. Nevertheless, the work presented along this thesis was conceived in order to deal with gene sets of any organism.

Chapter 2

Knowledge resources and functional annotation

In the middle of the 19th century, Gregor Mendel suggested the existence of discrete inheritable units that were later defined as *genes* in 1909 by Wilhelm Johannsen [HL15]. The genes are sequences of nucleotides that encode the synthesis of essential products of a particular organism. They became an important piece to understand the biological processes of a particular organism. For years, the improvements of techniques and tools to extract the information of a gene have allowed current machines to extract information of up to thousands of genes at the same time.

In this chapter, we present information that is available about genes and how it is represented. Then, we introduce the main strategies to extract information about gene sets, which involve genes that co-exist in similar experimental conditions, for improving their interpretations. This chapter is structured in the following four sections: section 2.1 relates a brief history of the origins of genes. In section 2.2, we describe the knowledge resources that store gene information and we present the limitations of studies with a single gene to understand complex biological contexts. section 2.3 shows the interest of studying a gene set in spite of a single gene and describes four main strategies to extract their information based on the annotation. At last, in section 2.4, we present the limitations and challenges of existing strategies.

2.1 Information supplied by a gene

In living organisms, the DNA (*DeoxyriboNucleic Acid*) is the genetic material that constitutes the genome. The DNA is formed of large chains made from four main nucleic acids: Adenine, Thymine, Cytosine and Guanine. The discovery of nucleic acids structure began in the 19th century after having isolated from the nucleus cells a component initially called nuclein [Dah05]. Throughout the 20th century, key studies have dealt with the identification of DNA as the genetic material [AMM44; HC52]. However, the real beginning of the modern molecular biology started with the works of Franklin and Gosling[FG53] with the extraction of `X-ray` diffraction image for the DNA and the double helix DNA structure proposition formulated by Watson and Crick[WC53]. After these contributions, the efforts were focused on the genome DNA study from a functional and structural point of view.

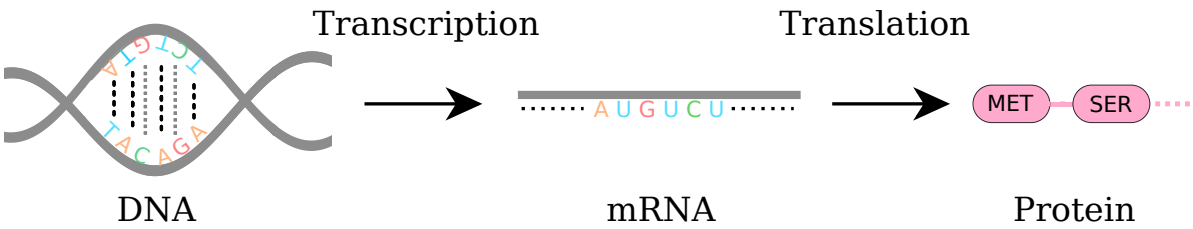


FIGURE 2.1: Simplified vision of the central dogma of molecular biology. A fragment of DNA is transcripted into *messenger* RNA (mRNA) where the difference with the DNA is the presence of the nucleic acid Uracil, instead of Thymine. Then, the mRNA is translated into proteins where each triplet of nucleic acids (codon) are translated into an amino acid (from a start codon to an end codon).

The genome DNA contains information about the organism. Thousands of DNA fragments (*genes*) must be processed to represent a phenomenon in the organism. Following the central dogma of molecular biology, the gene is transcribed into a *RiboNucleic Acid* (RNA) fragment, which is in turn translated into a protein (figure 2.1). Progress in understanding the procedures of genetic information transmission forced this initial fundamental scheme to be expanded in order to accommodate additional procedures that also happen, at least in some specific organisms. The information encoded in a DNA gene can be classified into two regions: *coding* and *non-coding*. The *coding* regions are the DNA portions of a gene that follow the commented dogma for producing a protein. They have an impact on an organism from a molecular or macromolecular point of view (known as *phenotype*). In contrast, the *non-coding* regions do not encode proteins. Some *non-coding* regions are transcripted into *non-coding* RNA such as transfer RNA, ribosomal RNA or regulatory RNA which are essential to the translation process for the *coding* regions [MEG06; Con+12; PD14].

The recent revolution in new sequencing technologies, as a part of the continuous process of adopting new innovative protocols [PK16] has strongly impacted our understanding of relations between *phenotype* and *genotype*. These experimental analyses are crucial to understand the gene features and these of their products as proteins. The extracted information related to a gene (structural or functional information) is stored under the form of annotations. The *gene annotation* is the connection between genes and the involved phenotype of the gene products. A gene or its corresponding products can be associated to different biological information including their localization in the genome, description about the variants or the products such as their expression in specific conditions. When the related information implies roles in which a gene or gene product is involved, this association is defined as a **functional annotation**. For a given gene, using annotation databases, one can expect to get information such as the biological functions, processes, pathways, diseases in which this gene is involved.

2.2 Knowledge resources providing biological annotation

To facilitate the access to this amount of information, the latter has to be stored and structured in systems that cover the different facets of biology. In the bioinformatics domain, such storage has mainly been made into relational databases during the last few decades [MBV12].

Nowadays, a particular attention is given to **ontology** for representing the knowledge necessary to describe biological information before storing it. The notion of ontology has initially been defined by Greek philosophers whose aim was to relate concepts to existence or reality [GOS09]. According to Uschold et al. [Usc+98], an ontology can take a wide range of forms, but necessarily includes a vocabulary of terms and some specifications of their meanings. In practice, an ontology can be simply represented as a set of **classes** and **properties** for describing objects in a given field of the real world [SGB00; HSG15]. These classes are then organized according to a graph structure thanks to hierarchical (taxonomic or meronymic) and associative (or transversal) **relations**. For example, in a biological ontology aiming to describe the different cells, their parts and their functions, *Eukaryotic_cell*, *Animal_cell*, *Mitosis* and *Meiosis* are classes. The *is_a* relation existing between *Animal_cell* and *Eukaryotic_cell* is a taxonomic relation and the *involved_in* relation between *Animal_cell* and *Meiosis* is an associative relation. For describing concepts and relations more comprehensively, formal ontologies are providing rules or constraints (called **axioms**) in a logical form through a formal language that allows to make deductions and potentially infer new knowledge. Such axioms enable to describe for example necessary conditions, like *each cell that is a eukaryotic cell must be involved in meiosis and mitosis*[1].

Ontologies have been used in a large number of disciplines for different purposes including legal information systems [BV97], geography science [FE99], semantic web [Din+07], biology or biomedicine [Don06; RSN08; RB11; HSG15]. The notion of biological ontologies or **bio-ontologies** has been introduced later. Their main application is the annotation and the integration of data [RB11]. Among the large number of bio-ontologies, the *Gene Ontology* (GO) is the most widely used [Con18] and is the central focus of the next subsection.

2.2.1 Gene Ontology

GO has been designed for describing the roles of gene products of any biological organism [Ash+00]. Unlike what is suggested by its name, GO originally corresponded to a controlled vocabulary rather than an ontology [SWS03; RSN08]. A controlled vocabulary provides a list of concepts, which regroup synonymous terms and are generally organized into a tree structure. More recently, efforts have been made for enriching GO by adding axioms to some of its concepts, converting it finally into a real ontology [Mun+11; The15].

GO structure

GO is organized according to three distinct categories, also known as ontologies, being *Molecular Function* (MF), *Cellular Component* (CC), and *Biological Process* (BP). MF describes the activities of gene products at a molecular level, CC presents the localization or cellular structure in which gene products perform their functions and BP includes the larger processes accomplished by multiple molecular activities. GO is composed of **classes**, **relations** and **axioms**.

Classes. The class, called "GO term" in GO parlance, represents a type of entity in a given domain. In GO, this entity thus corresponds to a function, localization or process of a gene

[1] Note that this is an example and there are more conditions to consider a cell as an eukaryotic cell

or gene product (*e.g.*, *protein transport*, *host cell membrane* or *cytokine-mediated signaling pathway*). Each term is associated with a unique identifier, such as `GO:0015031`, `GO:0033644` or `GO:0019221` respectively for the previous examples. Among GO terms, nearly 11,000 are molecular functions, little more than 4,000 are cellular components and almost 30,000 are biological processes.

GO terms are usually associated with additional information, including their definition, associated secondary or alternative identifiers, or a flag indicating if a given term is obsolete (figure 2.2). In addition, each GO term has one or more synonyms, such as *cytolysis* (`GO:0019835`) that has as synonyms *autolysin activity*, *necrosis* and *lysis*. The synonyms can be *exact*, *broader than*, *more precise* or *related* to the term in an imprecise way. Another interesting information is the *database cross-references* that connect the GO term to an identical or very similar class described in another knowledge resource. For example, the GO term *cytolysis* is linked to a class in UniprotKB-KW (`KW-0204`) and Wikipedia (`Cytolysis`).

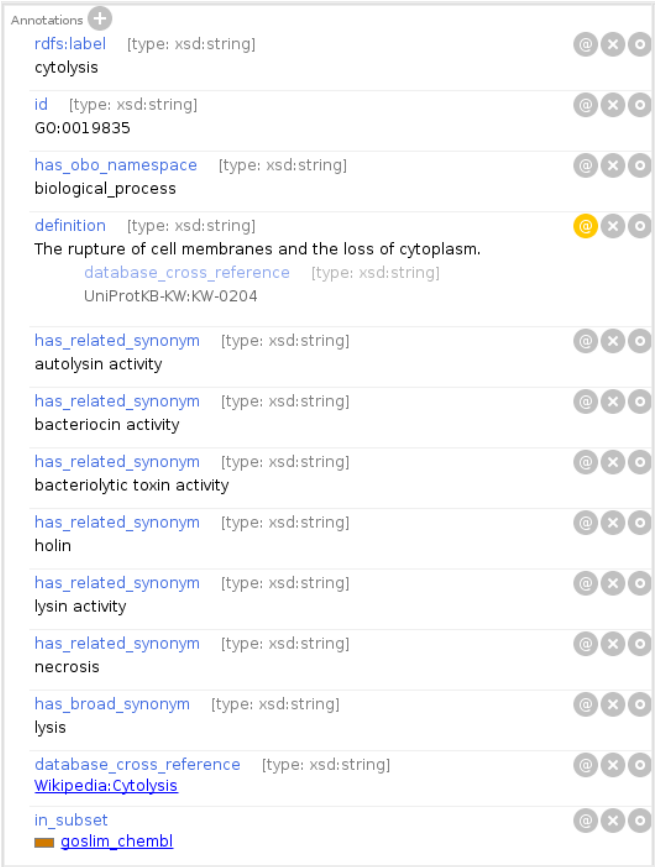


FIGURE 2.2: Screenshot from Protégé [Noy+03] showing information related to the GO term *cytolysis*

Relations. GO contains more than 45,000 terms connected through different kinds of relations: hierarchical (*e.g.*, *is_a*, *part_of*) and transversal (*e.g.*, *regulates* and more recently *starts_with*). These relations (corresponding to *edges*) in addition to the set of terms (corresponding to *nodes*) make the graph structure of GO be a *Directed Acyclic Graph* (DAG). A DAG is a graph that flows in one direction and does not generate cycles in their paths [Chr75; TS11]. The hierarchical relations in the DAG provide properties such as **reachability** or **transitivity**. These relations in GO are directed from the children to the parents. Moreover, in a DAG, a

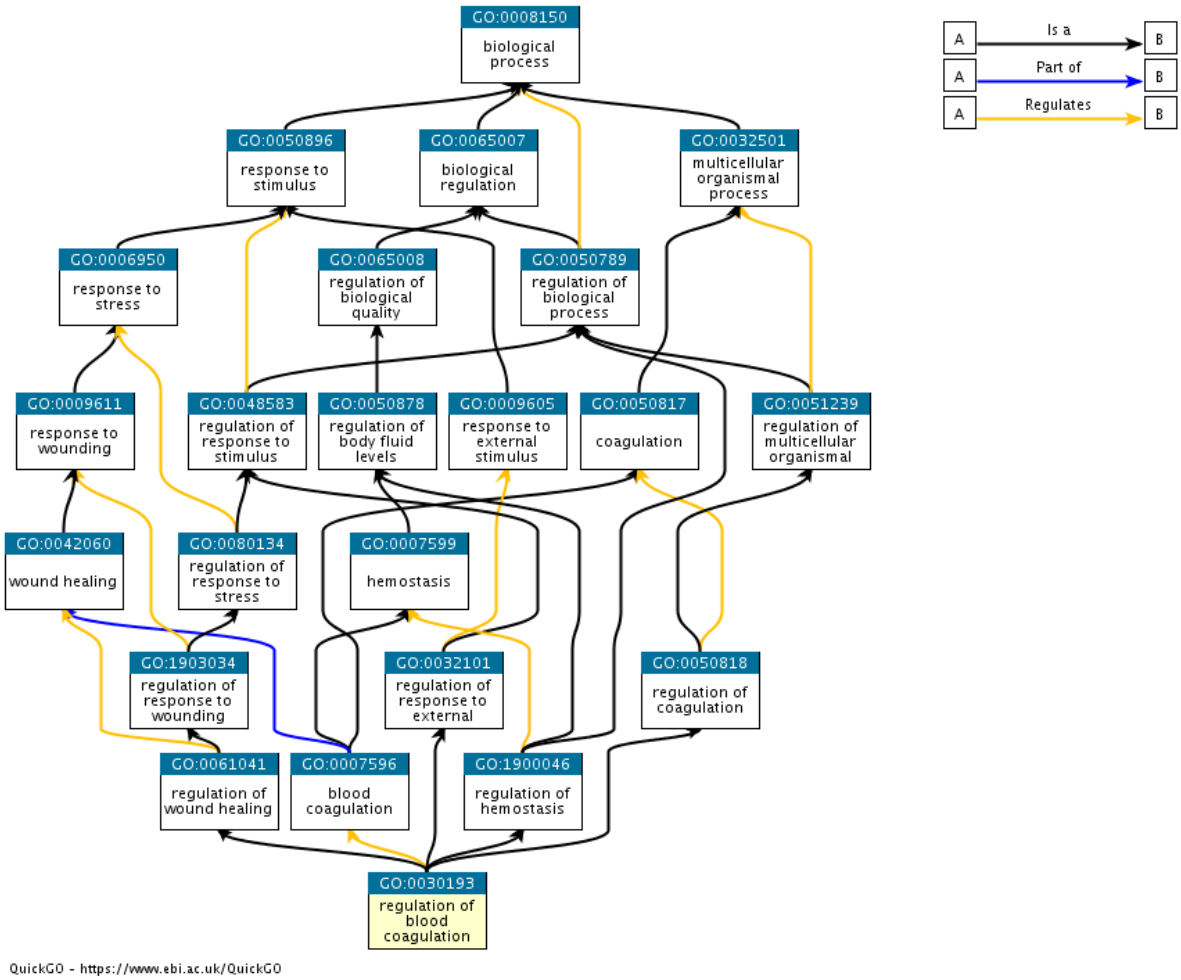


FIGURE 2.3: GO excerpt from the `QuickGO` website [Bin+09b]. This figure focuses on the ancestor terms of **regulation of blood coagulation** (`GO:0030193`). This sub-graph displays the three most important relations provided by GO.

node can be related with one or more parent *nodes*. **Reachability** is the ability to access a term from another term connected by a path. For example, a term t_1 is *reachable* from another term t_2 when there exists a path that starts at t_2 and ends at t_1. The **transitivity** in a DAG is whenever a term t_1 is related with a term t_2 and t_2 is related with a term t_3 then t_1 is related with t_3. Thus, if the relation between these terms is *is_a*, t_2 and t_3 correspond to the *ancestor* terms of t_1 while t_1 is a *descendant* term of t_2 and t_3. If the relation is direct (*i.e.*, only one *edge* separates terms), like between t_2 and t_1, the ancestor can be considered as a *parent* and the descendant as a *child*.

GO includes a considerable number of relations between terms, but the three main relations in GO are the following:

- **The *is_a* relation** constitutes the hierarchical taxonomy structure of GO. The transitivity feature of *is_a* allows to infer information between terms that are connected through these relations. Thus, the terms at the bottom of the hierarchy provide more specific information while the terms placed at the top of the hierarchy are more general. For example, the GO term *blood coagulation* (`GO:0007596`) is more specific than any of its two parents *hemostasis* (`GO:0007599`) and *coagulation* (`GO:0050817`), being themselves more specific than their parents, etc. (figure 2.3).

- **The *part_of* relation** is a hierarchical relation used to represent the part-whole relationship. Thanks to the transitivity feature, if a term A is part of a term B, the presence of the term A also implies the presence of the term B, while the inverse is not true. For example, in figure 2.3, *blood coagulation* has a *part_of* relation with *wound healing* (`GO:0042060`), meaning that if an event involves *blood coagulation*, we can affirm that a *wound healing* event is also involved. However, when we observe a *wound healing* event, the occurrence of *blood coagulation* cannot be affirmed.

- **The *regulates* relation** (and its sub-types *negatively_regulates* and *positively_regulates*) is a non-transitive transveral relation that is also important in the GO structure. This relation describes a specific case where one term is affected by the manifestation of another term, with an action of activation or suppression. Normally, the terms that are related to other terms according to a *regulates* relation are presented in the specific branch of *biological regulation* (figure 2.3).

GO format and tools. Many biological ontologies, such as GO, have been represented in the *Open Biomedical Ontologies* (OBO) format, because it has been specifically designed to implement biomedical ontologies. Simultaneously, GO has also been described according to semantic Web languages, and more specifically the *Web Ontology Language*[2] (OWL) [Has17].

There are numerous web tools that provide user-friendly front-end interfaces to explore and to investigate the GO structure [MC17] (*e.g.*, `AmiGO` [Car+09] or `QuickGO` [Bin+09b]).

GO is updated daily, but users can also edit it by using ontology editors like Protégé [Noy+03], OBO-EDIT [Day+07] or process it with libraries as OWL-API [BVL03]. These tools are interfaces allowing to explore, visualize and edit ontologies. They include a particular type of programs, called reasoners, that are run on the ontology to detect potential inconsistencies and to make deductions. Thus, a reasoner allows to classify new knowledge considering the structure of the ontology and the logical rules defined in the ontology [Abb12].

Gene Ontology Annotation

The roles of ontologies are multiple. Among them, the *annotation*, for which most bio-ontologies , including GO, were developed in the early 2000s [HSG15], is crucial. The principle of annotation is *to attach data to some other piece of data*, as defined by Oren et al. [Ore+06]. These authors have also reported that when knowledge from an ontology is used to annotate data, it corresponds to **ontological annotation**. The *GO Annotation* (GOA) database has been developed for this purpose, with the specific aim to associate GO terms to a gene product of a given organism in order to describe its biological roles [Cam+04]. The GOA is a consortium of different groups including UniProt [Apw+04; The16], *Mouse Genome Database* (MGD) [Bla+97; Bul+18], *Wormbase* [Gro+18], *Saccharomyces Genome Database* (SGD) [Che15], *Flybase* [MCG16], *dictyBase* [Kre+04; Chi+06] and *The Arabidopsis Information Resource* (TAIR) [Rhe+03; Lam+11]. Other groups out of the consortium such as EcoCyc [Kar+02] and the Functional Gene Annotation group at the University College London also contribute by providing annotations to enrich GOA. This assignation has been realized manually (when such an annotation was observed during laboratory experiments or according to bibliography) or automatically (*e.g.*,

[2] https://www.w3.org/TR/owl-features/

via the identification of sequence similarities). An **evidence code** is thus associated with each annotation for indicating its origin (*e.g.*, from which process and/or according to which source it was created) [RB09]. There are 22 different curated evidence codes involving *experimental, phylogenetic, computational* evidences as well as *author* and *curatorial statements*. The *experimental* evidence codes denote that the association between a gene and a GO term was provided by experimental methods. *Phylogenetically-inferred* annotations are derived from phylogenetic models where the loss or gain of a gene function is presented through a phylogenetic tree. The inference of this kind of annotation is supported by assertions produced by experimental analyses. The *computational* analysis evidence code corresponds to associations provided by *in silico* methods which use structural similarities like sequence similarities between two genes and inferring the information from the known gene to the unknown. *Author statement* provides associations by the interpretation of authors in the cited reference. The *curator statement* corresponds to any evidence that does not belong to any of the previous evidence codes. One such statement is the ND evidence code (*No biological Data available*) which means that the function, process or localization of a gene exists but the curator is not able to know what it is exactly. GOA also includes an electronic annotation evidence called *Inferred from Electronic Annotation* (IEA). IEA evidence corresponds to annotations that have not been manually reviewed even if the used method suggests quality assessments. Annotations with the IEA evidence are derived from two main pipelines: 1) manual mappings between external resources and GO terms, or 2) automatic transfer from orthologous gene products. Including annotations having the IEA evidence code within gene analysis has been discussed by several authors [Cam+05; Pes+09; Sch+09; PŠD11; Guz+12]. In GOA, 90% of gene-GO term associations have this IEA evidence code. Thus, considering this evidence code in the analysis might compromise results by the presence of false positives, whereas discarding them implies to ignore a large amount of potentially relevant information. For more details about this controversy, the reader may be interested in the work of Guzzi et al. [Guz+12] that summarizes the different assessments using IEA for analyzing different datasets and the conclusions drawn in each work.

In addition to evidence codes, **qualifiers** may be added to provide a precision about an annotation. A qualifier is a flag that modifies the interpretation of an annotation. In GO, there are three types of qualifiers: *NOT*, *contributes_to* and *colocalizes_with*. The most notable qualifier is NOT, which is used for stating explicitly that a gene product is not assigned to a given GO term.

The transitivity property of *is_a* and *part_of* relations in GO has a repercussion on annotations described in GOA, resulting from the *true-path-rule*. This means that if GOA contains an annotation between a given gene and a GO term, annotations can be deducted between this gene and the whole set of ancestors of the GO term [Hun+14]. Thus, the three root terms in the GO structure (BP, MF and CC) are related to any gene having at least one annotation in GOA with one of their descendant terms.

GOA format. The GOA is available in two plain text document formats: *Gene Product Association Data* (GPAD) and *Gene Association File* (GAF). GPAD provides information related to the annotation between the gene products and GO terms and it needs to be combined with its

companion file *Gene Products Information* (GPI) to extract the information of the gene products (*i.e.*, the gene id or the organism taxon). Instead, GAF includes all information provided by GPI and GPAD. It involves 17 columns describing both the association (GO term - gene product) and information regarding the gene product itself. A summary description of each GAF column is presented in figure 2.4 and detailed in Gaudet et al. [Gau+17] and on the GO website[3].

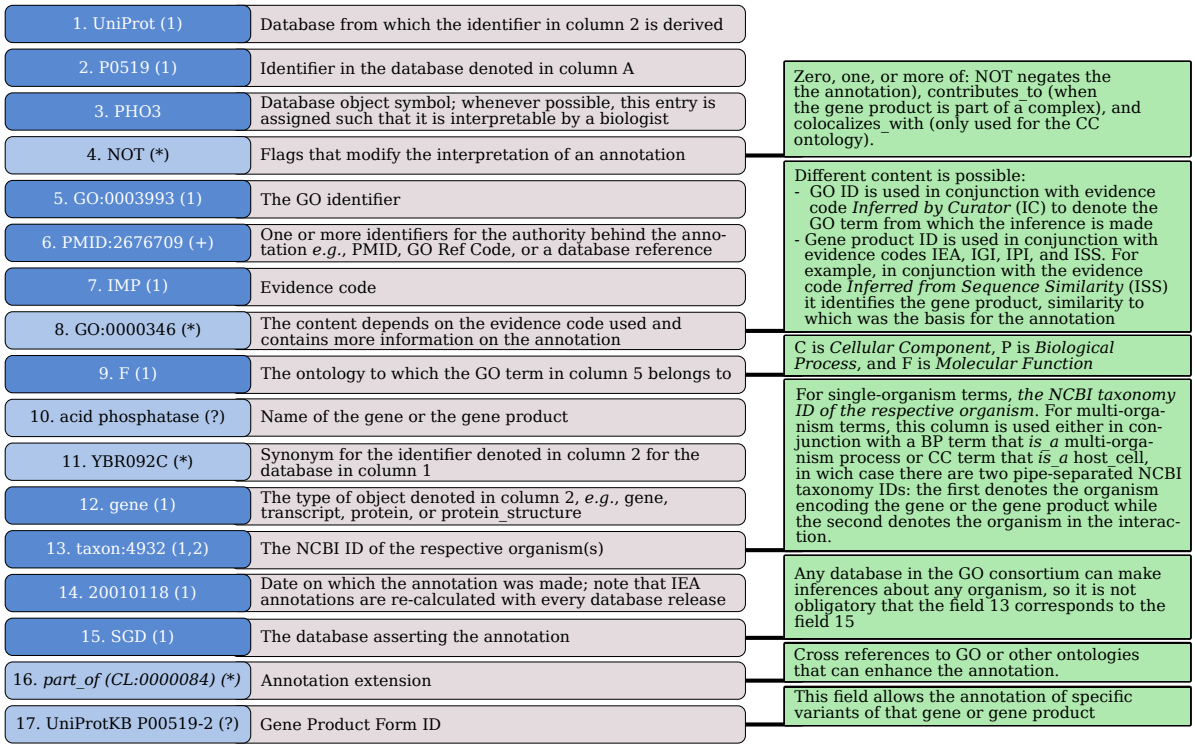


FIGURE 2.4: *Gene Association File* (GAF) 2.1 column description. Figure adapted from Gaudet et al. [Gau+17]. The *light blue color* indicates optional columns and the *green color* rectangles provide a more detailed description of certain columns.

2.2.2 Other knowledge resources

GO is not the unique knowledge resource that describes information about genes. Other databases and ontologies exist for describing other facets of the biological domain. As an illustration, the *National Center for Biotechnology Information* (NCBI) biomedical portal [OLe+15] provides easily accessible information and services related to biomolecules. Focusing on the functional annotation, a gene product can be involved in other situations than biological processes or functions or it can be placed in an organelle. Genes can also be the target of drugs, be expressed in a particular cell, be involved in important reactions or participate in pathways, be produced in specific conditions or phenotypes, or even be precursors of diseases. To illustrate that, we present in the following sections two examples of relevant knowledge resources to interpret biological data. The use we made of them is described in chapter 5, devoted to the integration of other knowledge resources within the `GSAn` framework.

[3] http://geneontology.org/docs/go-annotation-file-gaf-format-2.1/

Pathway sources

A biological pathway is a series of actions occurring in a cell to produce or catabolize metabolites. These actions can make interactions between molecules producing effects such as the regulation of gene expression, signal translation and metabolism. Structuring biological knowledge into pathways reduces the complexity of the numerous combinations of molecular entity interactions [Dom+18]. For example, *Pathguide* is a web server listing hundreds of biological pathway-related resources [BCS06]. Among the 702 knowledge resources listed in May 2019, 166 and 133 resources describe metabolic and signaling pathways, respectively. Moreover, some of these resources (*e.g.*, *Kyoto Encyclopedia of Genes and Genomes* or KEGG [KG00], *Reactome* [Jos+05], *Panther* [Tho+03], *Biocarta* [Nis01], and *WikiPathways* [Pic+08]) provide structured information describing pathways and their association with molecular entities as gene products.

Disease sources

A disease is defined as "*a verifiable, self-conscious sensation of dysfunction and/or distress that is felt to be limitless, menacing and aid-requiring*" [Kot80] and may refer to a set of genes as being the key actors. In genomics, two types of diseases are specifically studied: monogenic and polygenic. A monogenic disorder is produced by a defective single gene in the chromosome. This type of disorder is inherited according to Mendeley's rules, where depending on the dominance of a given gene, a gene could be expressed in a generation or another. The World Health Organisation estimates that over $10,000$ human diseases are monogenic, provoking a heavy loss of life. Examples of monogenic disorders are *Thalassaemia*, *Haemophilia*, *Fragile X syndrome* and *Huntington's disease*. Polygenic disorders, or non-communicable diseases, are usually involving multiple genes with complex interactions. This kind of diseases is not always acquired by parent heritage but is rather due to environmental factors (*e.g.*, chemical exposition, lifestyle). Some of the top 10 global causes of death in 2016 like *Diabetes mellitus*, *Stroke*, *Alzheimer disease* or some *Cancers* like *Trachea*, *Bronchus* or *Lung cancers* are polygenic disorders.

There are many resources describing human diseases and the extraction of complete information related to a specific disorder must be done by combining different disease resources. The most broadly used resource in the biomedical domain is the *Unified Medical Language System* (UMLS). UMLS is a compendium of many controlled vocabularies whose aim is to link health information or medical terms across different computer systems for enabling search engine retrieval, data mining, statistics or terminology research. The UMLS collects nearly 200 biomedical vocabularies. Some knowledge resources included in the UMLS are focused on diseases involving genes, such as the *Medical Subject Headings* (MeSH) [Lip00], the *Systematized Nomenclature Of Medicine–Clinical Terms* (SNOMED–CT) [Don06], the *Orphanet Rare Disease Ontology* (ORDO) [Vas+14], the *Disease Ontology* (DO) [Sch+11] or the *Human Phenotype Ontology* (HPO) [Rob+08].

2.2.3 Limitation in studying single genes

Knowledge resources, such as GO, are useful to determine the functions of an annotated gene (or a gene product) in a given organism to understand its roles and implications in biological

contexts. Nevertheless, the information provided by a single gene is not enough to understand complex biological contexts. Considering a whole gene set to study a biological process, a pathway or a complex disease is more useful than considering genes individually [Sch+12]. For example, in clinical applications, until recently, single genes were used in genetic tests to confirm a clinical diagnostic such as cancer, neurological disease or heart disease. Even if they provided good results for specific monogenic diseases, most of the diseases are not induced by the mutation of a particular gene but rather involve multiple genes [Nev+13]. Currently, studies consider groups of genes that are often related as they co-occur in the same pathway [Sub+05]. As mentioned in section 1.2, several genes that are slightly differentially expressed are more relevant than only one gene with a strongly different expression pattern. Thus, the next section describes the challenges raised by the interpretation of gene sets to better understand a biological context.

2.3 Interpreting gene sets

Studying a single gene by exploiting its annotation corresponds to traditional approaches in biological research [HSL09]. Today, with the new sequencing technologies, the use of gene set approaches has become increasingly popular. Over the years, many works have been dedicated to create, manage and curate gene sets [Ant11; Bak+11; Cul+11; Li+13; CB14; Lib14]. *GeneWeaver* is a vast gene set repository for storing, curating and analyzing gene sets from heterogeneous data sources [Bak+15]. Currently, *GeneWeaver* includes more than $199,650$ gene sets largely derived from high-throughput expression techniques, mutation and perturbation screens, and curated biological relationships. The *Molecular Signatures DataBase* (MSigDB) is another gene set repository widely used for the functional annotation of gene sets [Lib+11]. In the 6.8 version, MSigDB includes $17,810$ gene sets divided into eight major collections, and several sub-collections derived from specialized resources such as GO, KEGG, *TRANScription FACtor database* (TRANSFAC) [Win+96] and L2L [NW05]. A recent functional annotation method, called `EGSEA`, provided a gene set database of $25,000$ gene sets involved in 16 collections, where the gene sets of MSigDB are included into a single collection [Alh+16].

This new field of research has become unavoidable over the last two decades and it is based on the inference of gene sets according to a comparison of experimental results under diverse conditions. In such a context, sets of genes that are contributing to specific phenotypes are inferred based on various experimental conditions [Cha+08; Bin+13; Li+13; CB14]. For example, Chaussabel and Baldwin [CB14] used experimental conditions related to various diseases to decipher the genes that may be involved in the innate or specific immune response. An additional issue is the interpretation of these gene sets using the information available for each gene, which is based on annotation terms derived from a wide range of resources. The functional interpretation of genes is decisive for the understanding of life and it involves to make use of the whole subset of terms that can be related to these genes. With this goal, public knowledge resources such as KEGG or GO are generally used to recover information which is already available about the genes involved in these sets. However, the number of terms in such subsets can be relatively large. For example, considering that human genes are annotated on average with 10 terms, a subset of 100 genes may lead to several hundreds of terms including an important and useless overlap of similar terms. Moreover, when studying several heterogeneous gene sets, the number of terms increases, thereby involving thousands of

them. Thus, the manual expertise to clearly decipher the main functions that may be related to the studied gene set(s) is time-consuming and becomes impracticable when the number of gene sets increases, as it is the case in vaccine/drug trials. This way, even for analyzing a tiny number of genes, scientists may be overwhelmed with the large amount of information accessible to the community. Thus, automatic methods have been proposed to facilitate the analysis of gene sets.

In this section, we define four areas of studies that help understand the biological function of a gene set. First, we present existing enrichment methods for annotating a given gene set. We then define some alternatives based on gene annotation and ontological characteristics. Thirdly, we describe the use of visualizations to facilitate the exploration and interpretation of gene set annotation. Finally, we present some attempts to integrate biological knowledge resources for providing a wider biological context.

2.3.1 Gene set annotation by classical enrichment techniques

Attributing a biological function to a given gene set first requires to analyze the information given by annotation terms associated with each gene of this set. A key issue of gene set annotation is the too high and too heterogeneous amount of annotation terms that makes their global analysis difficult. Decreasing the number of terms while keeping the most informative ones is a major challenge to understand the biological implications of a gene set [BPG16]. Indeed, these informative terms should be representative of the initial big amount of annotation without redundancy between them, while providing the essential information. One popular approach used to interpret biological information related to gene sets is based on enrichment statistical methods. The underlying idea of these methods is to compare, on the basis of their annotations: (i) genes sets (which are co-expressed or co-regulated in a given phenotypical condition, for example), and (ii) a reference (*e.g.*, the complete genome or a gene set generated randomly whose size is comparable). During the last decade, many statistical enrichment methods aiming to decipher gene sets and to understand their biological meaning have been developed (for a review, see [HSL09]).

Enrichment methods have been divided into three categories or classes [HSL09]: *Singular Enrichment Analysis* (SEA), *Gene Set Enrichment Analysis* (GSEA), and *Modular Enrichment Analysis* (MEA). Later, Khatri et al. [KSB12] proposed the following alternative classification of enrichment methods by focusing more on pathways' application: *Over-Representation Analysis* (ORA), *Functional Class Scoring* (FCS), and *Pathway Topology* (PT)-based approaches. These classifications are widely accepted and have been popularized by many authors as [TH10; Nun+14; Yan+17; Fab+19]. By looking at the definition of each different category, equivalences can be established. The SEA and GSEA classes are respectively equivalent to ORA and FCS, while MEA and PT are slightly different in their descriptions.

The different categories are described below with an illustration of each of these categories in figure 2.5.

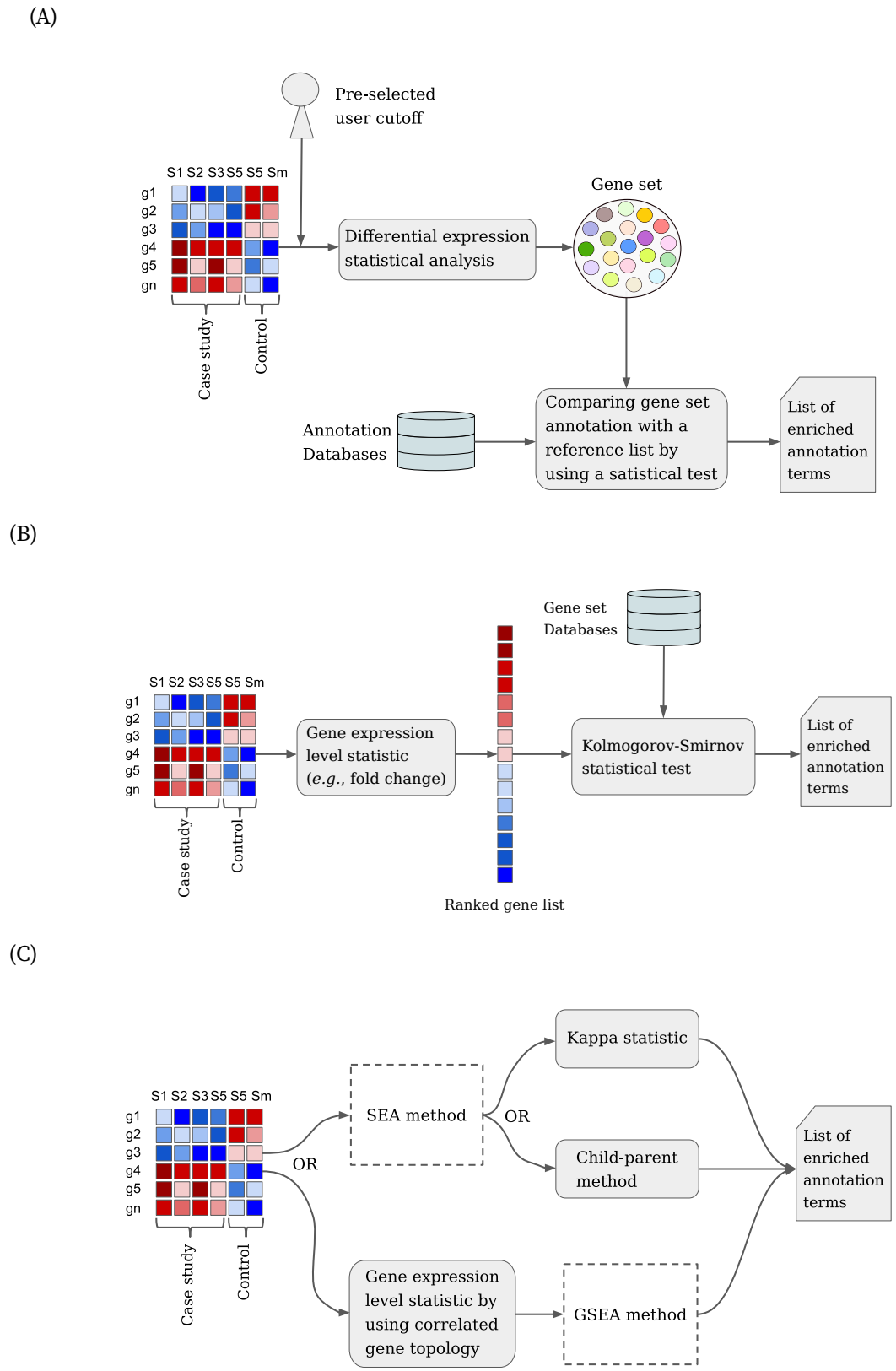


FIGURE 2.5: Enrichment analysis schemas for the three classes: (A) SEA or ORA, (B) GSEA or FCS, and (C) MEA and its sub-category PT. Each schema represents the procedure, using a gene expression matrix as input and generating a list of enriched terms as output.

Class I: SEA or ORA

These methods evaluate the statistical significance (a result statistically significant means that it is highly unlikely that this result was produced by chance) of a fraction of genes for a particular annotation term. Using exclusively a gene list, each associated annotation term is compared to a reference gene list using a statistical test to compute the enrichment *p-value* (figure 2.5A). One term is enriched when it exceeds a threshold predefined by users. Statistical tests that can be used to determine the enrichment *p-value* include the Chi-square, the Fisher's exact test, the binomial probability, and the hypergeometric distribution (see a description of them in [Riv+06]). Thus, a list of enrichment terms is proposed to summarize the biological function of the preselected gene set. Examples of tools from this class are `GO-stat` [BS04], `MappFinder` [Don+03], `KENeV` [Pil+15], `ClueGO` [Bin+09a], `g:Profiler` [Rei+07; Rau+19], and `DAVID` [Den+03; Hua+07].

Class II: GSEA or FCS

These methods are suitable to compare two biological studies (*e.g.*, tumor cells versus normal cells) [HSL09; TH10]. An advantage of the GSEA method is that it does not require a predefined threshold to get interesting genes. Instead, GSEA uses all genes for the analysis. Generally, the enrichment process performed by these methods can be divided into three steps [AS09; KSB12] (figure 2.5B). First, a gene-level statistical approach (as t-test, Q-statistic or z-score) computes the differential expression of each gene and the genes are then ranked by their degree of differential expression or *fold change* [TH10; Fab+19]. Secondly, an enrichment score distribution is calculated by using the ranked gene list for a particular annotation category whose prior knowledge is available (*e.g.*, a given GO term). A *p-value* is determined by using statistical tests, such as the Kolmogorov–Smirnov test [Sub+05]. Tools falling into the GSEA class include the `GSEA tool` [Sub+05], `EGSEA` [Alh+16], `sigPathway` [LTP13], and `PCOT2` [SB10].

Class III: MEA or PT

The description of this last class is slightly different due to the fact that MEA and PT are not equivalent. MEA was defined by inheriting the basis of SEA methods including an extra term-to-term network discovery algorithm [HSL09], while the PT method is essentially similar to the GSEA method except that it uses the pathway topology to compute the gene-level statistical value (figure 2.5C). The additional network algorithm uses the hierarchical structure of a knowledge resource (generally GO) or the correlation between genes. Tools like `DAVID` [Hua+07], `Ontologizer` [Bau+08], `topGO` [AR09], `GeneCodis` [TNP12] and `EnrichNet` [Gla+12] perform a network algorithm, such as a *parent–child* method or the *kappa statistic* after the enrichment analysis computed as in the SEA class (figure 2.5C). Focusing on pathways, `ScorePAGE` [Rah+04] and `Pathway-Express` [Dra+07] are examples of enrichment analysis tools that apply topological similarities to obtain the gene-level statistical value (figure 2.5C).

2.3.2 Opportunities of ontological annotation

As commented in section 2.1, information regarding genes or gene products can be stored in knowledge resources, as bio-ontologies, for describing different phenotypes. The advantage of using an ontology is that all the knowledge it describes as well as its structure can be

used to have additional information regarding annotation terms [LJM05]. In particular, each term in GO has many features which can be used to determine its similarity with other terms.

Another solution, also based on GO and currently used along with enrichment analysis, is called the GO subset annotation [Con04a; Rhe+08; Con18]. These techniques reduce the complexity of the ontology by selecting only some terms specific to a given domain [Rhe+08; DSR10; HDA11; Pri+13; ZYL17].

We first introduce the features that may characterize each term in GO. Then, we present semantic similarity measures that can be used to compare two terms by using these features. At last, we describe the approach consisting in the creation of a subset of GO terms for realizing analyses.

Features of a term

The GO structure is defined as a DAG, in which terms are connected through direct links. This type of structure provides characteristics to GO terms, such as the depth (*i.e*, the longest path from the root of the ontology to the term), the set of its descendants (*i.e*, its child terms and recursively until the leaves), the set of its ancestors (*i.e*, its parent terms and recursively until the root), and its *Information Content* (also referred as *IC*). The *IC* is a quantitative value that determines the information carried by a given term. Mazandu and Mulder [MM13] classified the different types of *IC* into two families: *annotation-based* and *topology-based*. The annotation-based family computes the *IC* by using information from external resources [Res95] while the topological-based family uses information from the structure of the ontology in which the term is defined [SVH04; Wan+07; ZWG08a; SA10; SBI11; MM12a].

Semantic similarity and relatedness

The identification of similarity between two terms may be crucial for facilitating the interpretation of gene set annotation. Considering terms associated with the genes of a given set, their similarity can be used to group these terms into categories with the aim to minimize the number of terms describing the biological context of this gene set. The semantic relatedness and semantic similarity are measures assessing the resemblance in meaning of two terms within an ontology.

According to Pedersen et al. [PPM04] and Pesquita [Pes07], semantic relatedness and semantic similarity are two related, but distinct notions. Thus, semantic relatedness is a broader notion making use of several relations between two concepts (*e.g.*, *is_a*, *part_of*, *regulates*) while semantic similarity is a special case of relatedness that makes only use of taxonomic relations. Nowadays, the number of proposed measures of semantic relatedness and similarity is extensive [PPM04; Pes+09; Guz+12; MCM17]. Pedersen et al. [PPM04] proposed a classification of semantic relatedness and similarity in biomedical knowledge according to three types: path finding, *IC* or context vector. Pesquita et al. [Pes+09] proposed a survey of semantic similarity measures used for comparing terms in the context of GO. The authors categorized these measures according to the three following classes:

- *node-based* measures which make use of features of GO terms,

- *edge-based* measures which leverage relations that exist between GO terms,
- *hybrid* measures which mix methods from the two previous classes.

[Guz+12] also did a review of existing semantic similarity measures. Authors displayed the latter within a Venn diagram, which facilitates the identification of features that are used by many measures but also strategies that exploit only a few of them. In addition, measures which make use of multiple features can also be easily identified, being at the intersection of multiple ovals. Mazandu et al. [MCM17] recently presented an additional review of semantic similarity measures (referred to in their paper as "term semantic similarity"). These authors provided an exhaustive list of the different *ICs* that have been proposed in the literature and refined the classification proposed by Pesquita et al. [Pes+09] by adding the following subcategory to the *node-based* measures: *graph-based* measures. This subcategory has been previously introduced in Mazandu and Mulder [MM12a] for specifying measures based on the ancestors and/or descendants of the terms to be compared.

Semantic similarity measures are used in a wide range of applications and domains. Harispe et al. [Har+15] described three main domains of application: *Natural Language Processing, Knowledge engineering, semantic web and linked data* and *Biomedical informatics and bioinformatics*. Focusing on *Biomedical informatics and bioinformatics*, some works have presented the interest to use semantic similarity measures to evaluate the functional similarity between genes [Frö+07; Wan+07; Du+09; Yu+10], to study protein-protein interactions [XDZ08], to predict gene annotation [Ye+05; MM12b] or cellular location [LD06], to assess gene set coherence [RRN09; DA11], to compare gene sets [BDD14], and to improve the gene set annotation [Xu+09; Sup+11; Yu18].

Many tools making use of semantic similarity measures have been proposed for performing a gene set annotation based on GO. Supek et al. [Sup+11] included in the `REVIGO` tool an *a posteriori* analysis of the terms computed by enrichment methods by using semantic similarity to reduce redundant information. Other tools such as `FunSimMat` [SA07], `G-SESAME` [Du+09], `GFSAT` [Xu+13], `DAGO-Fun` [Maz+15], `GOGO` [ZW18], `GOSim` [Frö+07] and `GOSemSim` [Yu+10] compute semantic similarity and/or relatedness between two or more GO terms as well as functional similarity between genes. `FunSimMat` offers pre-computed functional similarity values of proteins and protein families. Also, this tool allows to compute semantic similarities from a list of GO terms compared *all-against-all* [SA07]. `G-SESAME` is a set of tools providing semantic and functional similarities by using classical approaches and their proposed measures [Wan+07; Son+14]. These tools compute similarity between two genes or proteins as well as two sets of GO terms. Moreover, `G-SESAME` provides a function called "knowledge discovery" that groups genes by using a clustering method according to their functional similarity. `GFSAT`, `DAGO-Fun`, and `GOGO` provide similar functionalities like `G-SESAME`. Additionally, `DAGO-Fun` includes a gene set enrichment tool (part of the SEA class) with an *a posteriori* analysis that applies a semantic similarity measure. At last, the R packages `GOSim` and `GOSemSim` allow to compute semantic and functional similarity into a script. An advantage of these tools compared to stand-alone or web tools is the possibility to choose the annotation and the ontology version in an R program for computing similarity. Moreover, a Java library (not being GO-specific), called *Semantic Measures Library* (`SML`), has been developed and involves different semantic similarity and relatedness measures [Har+13]. `SML` includes more than 50

semantic and functional similarity measures and can be used with different ontology formats.

Other tools use semantic similarity measures combined with graph theoretical approaches to annotate gene sets. Lee et al. [LHK04] described an approach to annotate gene sets by using the structure of GO (transformed into a tree structure[4]) by using semantic similarity measures. Authors did not mention *semantic similarity measures* but their distance approaches (*i.e.*, *maxPd* and *avgPd*) are actually based on the lowest depth of the common ancestor between two GO terms (called *Lowest Common Ancestor* or LCA), which is a feature used to compute some semantic similarity measures.

GO subsets

GO subsets (or GO slims) provide a overview of biological processes or functions for a single organism, clade of organisms or a broader biological area [HDA11; Pri+13]. Thus, very specific GO terms can be discarded to provide a finer granularity of information. Moreover, some complete branches may be irrelevant for a particular organism or domain. GO subsets aim at selecting only relevant GO terms for a given purpose. For example, the GO terms *mitotic cell cycle* (`GO:0000278`) and *meiotic cell cycle* (`GO:0051321`) have to be removed if one tries to create a GO slim dedicated to the description of bacterial organisms.

Eleven customized GO slim annotations currently maintained by the GO consortium mainly focused on various subgroups of organisms (*e.g.*, plant, yeast). Also, tools like `AmiGO GO Slimmer` [Car+09], `SGD GO Slim mapper` [SH11], `Map2SLIM`[5] or `QuickGO` [Bin+09b] allow users to create their own GO slim. While many uses of GO slim annotations have been proposed for specific projects [Arn+09; GMR10; Pri+13; Woo+19], only a few attempts have been made to automatically compute them [DSR10; JL10]. Davis et al. [DSR10] calculated the optimal reduced GO graph by using graph and information theories. Jin and Lu [JL10] identified informative subsets of GO terms keeping the maximal semantic information by using the frequency of annotation terms. In contrast, the number of methods related to their exploitation is more sizable as they mainly depend on the use of various similarity measures to compute proximity between annotated GO terms and slim customized annotations. Put simply, genes are assigned to a specific category given by the slim annotations [NSG14]. These methods may also be combined with enrichment analyses where the set of annotations is, from the beginning of the analysis, reduced to a subset of terms [PGC09; DSR10; GMR10; Cou+16; SVW16].

2.3.3 Summarizing results with visualization

Visualization is very useful to explore scientific knowledge. The adage *a picture is worth a thousand words* refers to the notion of representing complex ideas into a single picture. In the same way, a visual metaphor is worth a thousand data. Thus, the increased amount of data generated in the field of genomics quickly suggested the need to provide visualization methods [Hel+98].

[4] A *tree* is a DAG where each child node has one and only one parent node
[5] ttp://search.cpan.org/ cmungall/go-perl/scripts/map2slim

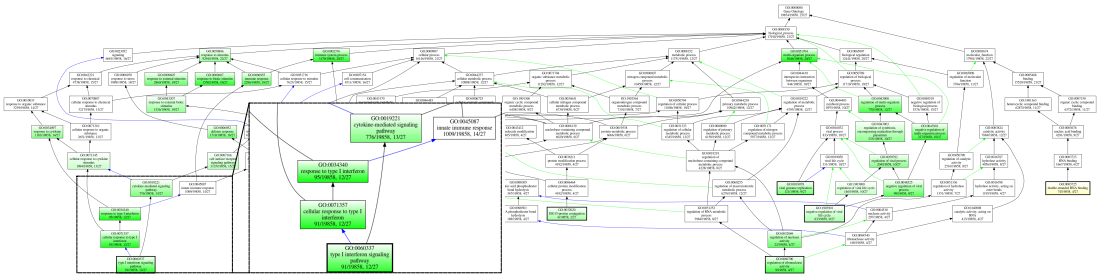
The number of visualization techniques in the biological domain considerably increased in the last 15 years [Ker+17]. This increase is also associated with a high diversity of visualizations for the same dataset. Over the recent years, many surveys have been realized to report different techniques such as text visualization [KK15], graph or tree visualization [Sch11; Bec+14; VBW15], multifaceted visualization [KH12]. Lately, an online tool, called *BioVis Explorer*, was introduced with the aim to collect a large number of visualization papers dedicated to biological data [Ker+17]. With this tool, authors aimed to reduce the complexity of the large variety of visualizations to deal with biological data. To do so, they computed dissimilarity measures according to several attributes such as the type of data or the task to be solved. Then, using a technique of *MultiDimensional Scaling* (MDS), authors depicted an interactive visualization involving all the works related to the biological domain that they have reviewed.

Mougin et al. [Mou+18] reported existing approaches implemented to visualize clinical and omics data by considering different attributes and visual metaphors. Authors categorized the different visual metaphors according to the data dimensionality and related to various analysis tasks:

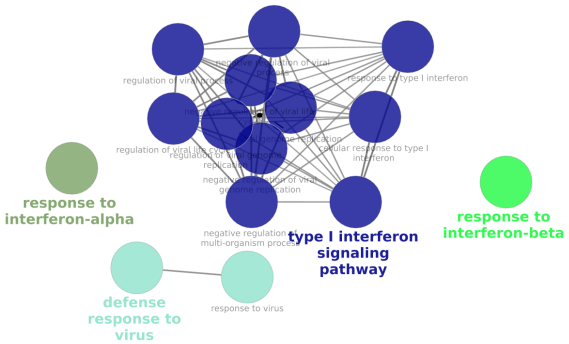
1. **Data dimensionality**
   - *1D data.* The main techniques used for visualizing this type of data are *bar charts* and *circular* or *pie charts*.
   - *2D data.* Two popular techniques for this category are *2D scatter plots* and *heatmaps*.
   - *nD data.* Very few techniques allow to represent multidimensional data. Techniques such as *heatmaps* including different attributes (*e.g.*, position, color) and *parallel coordinates plots* are examples of these techniques. Normally, these visualizations display their attributes using $1D$ or $2D$ representations. It is also frequent to use statistical methods to reduce the dimensionality (to $2D$ or $3D$) and then to represent data with metaphors such as scatter plots.
2. **Related analysis task**
   - *Relational data.* The main data falling under this category are graphs or networks. The following two techniques are mainly used for visualizing this category: *node-link* and *matrix-based* diagrams.
   - *Stamped-data.* The visualization techniques focused on temporal-stamped are represented with *line charts* or *timelines*.

Focusing on gene set annotation, specially on GO annotation, some tools propose enhanced visualizations which can be mainly classified into five categories [SŠ17]: *node-link diagrams*, *treemaps*, *semantic similarity spaces*, *heatmaps*, and *word clouds*. In *node-link diagrams*, some methods first extract the computed annotation terms and their relations within the GO structure and then draw a DAG using a layered/hierarchical drawing algorithm [ZKS05; Car+09; Ede+09]. Other methods (*e.g.*, Supek et al. [Sup+11]) connect together annotation terms according to their semantic meaning (computed using the ontology) and then use a force directed algorithm to take into account the similarity score for drawing the graph. Methods

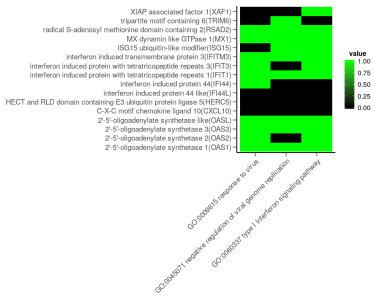
(A)



(B)

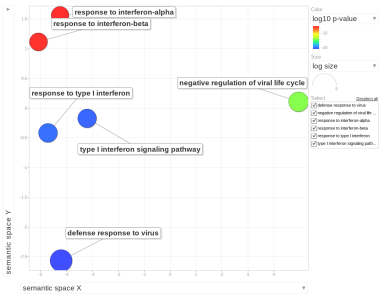


(C)



(D)

(E)



(F)

FIGURE 2.6: Examples of visualization results proposed by different enrichment tools. By using a gene set involved in the interferon signaling pathway [CB14], the visualization results are presented in several forms. The **node-link diagram** is shown in (A) using an extract of GO in which the over-represented terms are colored in green (obtained from `Ontologizer` [Bau+08]) or (B) creating new links according to the similarity of terms based on gene correlation (from `ClueGO` [Bin+09a]). (C) A **heatmap** represents the adjacency matrix of genes and GO terms (from `DAVID` [Hua+07]). Using the *a posteriori* analysis of `REVIGO` [Sup+11], (D) a **treemap** groups terms that are semantically similar, (E) a **semantic similarity space** is represented by a scatter plot, and (F) a **wordcloud** representation displays GO terms or associated keywords. The font size and colors are related to the *p-value* provided by the enrichment analysis.

from the *treemaps* category use space-filling visualization techniques to render the DAG (computed using similar strategies). For instance, Supek et al. [Sup+11] use treemaps that represent only two levels of the hierarchy and avoid the duplication of annotation terms by making use of the computed p-values. The *semantic similarity space* category uses principal component analysis (PCA) or MDS to represent annotation distances in spaces of 2 or 3 dimensions by using scatter plots [Sup+11]. The *heatmaps* category proposes a binary association between genes and annotation terms [Den+03] or groups similar annotation terms [Zee+11] in a matrix-based diagram where a row represents a gene and a column represents an annotation term or a group of terms. Finally, *wordclouds* can display the name of GO terms whose font size is related to the importance of the term (established according to p-values computed by enrichment analysis or to the *IC* value) [Sup+11; Nin+18]. figure 2.6 shows examples of metaphors from each of the **five categories** applied to a gene set that has been manually annotated as interferon [CB14]. This gene set was analyzed by using enrichment analysis computed by `Ontologizer` [Bau+08] (figure 2.6A), `ClueGO` [Bin+09a] (figure 2.6B), `DAVID` [Hua+07] (figure 2.6C) and the *a posteriori* analysis of `REVIGO` [Sup+11] (figures 2.6D to 2.6F)

These tools are only some examples of the existing tools dedicated to the visualization in the biology domain. Pavlopoulos et al. [Pav+15] proposed a catalog of web and stand-alone visualization tools developed in the biology field. This catalog represent the programming language libraries and visualization tools proposed until 2015 for biology networks, pathways, genome alignments, genome browsers, comparative genomics, phylogeny tree viewers and microarray and `RNAseq` analysis viewers.

2.3.4 Using multiple knowledge resources to improve the biological understanding

Despite the effectiveness of data generation from the massive outburst in technological advances for generating and processing large biological data, genomics, transcriptomics and proteomics are still separate fields of research [Man+16]. Most knowledge resources can only answer specific biological questions. For example, GO describes processes, functions or localizations of gene products while *Reactome* has been created to model the interaction of these products within pathways. Moreover, some knowledge resources may describe similar (or same) notions, thus generating duplication and potentially resulting in redundancy [KPL03]. For example, the knowledge resources KEGG [KG00], *Reactome* [Jos+05] and *WikiPathways* [Pic+08] all contain pathway information from different semantic spaces. It is then critical to integrate overlapping, but also complementary, knowledge resources in order to unify information and to be able to provide a global view on omics data.

According to Keet [Kee04], the notion of **integration** means *anything ranging from integration, merge, use, mapping, extending, approximation, unified views and more*. Integrating two or more knowledge resources is thus a challenging task. In this frame, Hernandez and Kambhampati [HK04] described challenging characteristics, such as the *variety of data* or *representational heterogeneity*. Recently, Mısırlı et al. [Mıs+16] claimed that the lack of acceptance of a standard format was *one of the major problems in data integration*. The W3C[6] offers

[6] https://www.w3.org/

an extensible base, the *Resource Description Framework* (RDF), for describing knowledge resources [LSW98]. This language represents knowledge according to triplets (*i.e.*, subject, predicate, object). An extension of RDF is OWL, which enables to describe ontologies such as GO [Bec+04]. Other standard formats have been widely used in the biological domain such as *Biological PAthway eXchange* (BioPAX) [BCS04], *System Biology Markup Language* (SMBL) [Huc+03], OBO [Day+07] and *Biological Expression Language* (BEL) [HDH18]. BioPAX was conceptualized for representing pathway knowledge [BCS04]. The aim of BioPAX is to be able to integrate, exchange, visualize and analyze biological pathway data. Pathway-based knowledge resources like *Reactome* and *WikiPathways* are available in this format. SMBL represents and exchanges models between simulation tools. It represents chemical reactions and is frequently used in system biology [Huc+03]. The OBO Foundry is a consortium whose aim is to integrate interoperable and well formed ontologies. The ontologies in OBO Foundry such as GO, *Sequence Ontology* (SO) or HPO, are described both in OWL and OBO formats. The OBO format is an alternative to OWL for describing biological ontologies and it is designed to be human readable and editable. At last, the BEL has been proposed as a robust format integrating information from multiple biological domains [Kha+15; NKH15; Emo+17; Iya+17; Hoy+19]. BEL is a language for representing scientific observations in the life sciences. BEL relates terms that denote biological entities (*e.g.*, genes, messenger RNAs, diseases and drug compounds) and biological processes (*e.g.*, tissue damage, cell cycle and kinase activity) based on *subject–predicate–object* triples [Sla14]. The triples in BEL are presented according to statements where each triple describes a scientific finding. Thus, BEL allows to build biological models (provided by ontologies, databases or thesauri) with experimental results in a semantic way [Sla14].

All these standard formats have been developed to facilitate the integration of different resources by reconciling the same or equivalent information. However, the existence of different formats makes arduous the integration of different resources. An even more challenging issue is the heterogeneity in the way similar entities are described. For that, many efforts have been lead on ontology-based integration, also known as ontology mapping or ontology alignment, that uses methods or techniques to establish semantic links (or *mappings*) between entities (*i.e.*, classes, relations, instances) of different ontologies or sub-parts of a single ontology. For years, many matching approaches have been proposed and classified into several categories, such as *string-based*, *language-based*, *graph-based*, *instance-based* (for details of each category and different tools, see Otero-Cerdeira et al. [ORG15] and Shvaiko and Euzenat [SE13]). Other strategies based on data mining aimed to infer mappings between resources. For example, The `HPO2GO` mappings that relate HPO and GO terms make easier the understanding of the origin of a phenotype produced by the loss or alteration of one or more gene functions [Doğ18]. `INTEGRO` centralizes disease annotations from different knowledge resources into DO [CGV18]. By applying methods based on association rules, pertinent relations may be inferred between two different ontologies or two sub-parts of a single ontology (for a review of such methods, see [Nau+13] and [GMC14]). Faria et al. [Far+12] applied association rules mining focused on the consistency of annotations. Benites et al. [BSS14] used a data mining approach based on rare associations to discover new relations between different knowledge resources or different parts from a single ontology (*i.e.*, among different sub-ontologies). Manda et al. [MMB13] presented a data mining approach called *Multi-ontology*

data mining at All Levels (MOAL). MOAL mine the different sub-ontologies of GO in order to enrich them generating GO annotation candidates and generating new relationships between terms.

Initiative projects, like Bio2RDF or syBioOnt, provide to the bioinformatics community a knowledge resource integrating different information from genes like their functions, interactions with other genes, drugs or their disease phenotype [Bel+08]. Bio2RDF designed some rules in a semantic model by converting every involved knowledge resources into the RDF format, which thus facilitates their integration. Recently, syBioOnt was designed to integrate and to formalize the representation of different knowledge resources into an ontology defined in OWL [Mıs+16].

2.4 Challenges related to the annotation of gene sets

Throughout the previous section, we presented techniques focused on the annotation of gene sets. Nevertheless, each solution exhibits remaining challenges.

Enrichment analysis. These methods aim to retain the over-represented terms (*i.e.*, overused by the genes in the gene set) without considering the relevance of the information and the specificity of these terms. Thus, the results supplied by enrichment tools consist of lists of numerous over-represented terms, and an *a posteriori* stage remains necessary to remove the potentially redundant information. In this frame, MEA tools like DAVID [Jia+12] use an *a posteriori* analysis of the annotation terms co-utilized by the genes to cluster the genes into potential groups of similar information. However, manual expertise still remains crucial and becomes unachievable if the number of gene sets to be analyzed is too high. Moreover, these methods provide redundant information by selecting terms that are hierarchically related, leading to difficulty and occasional bias in correctly interpreting the results. Moreover, these methods tend to focus on the most studied genes and provide gene set annotation results that cover a limited number of annotated genes [BLG15; HTK18; Tom+18].

Knowledge-based solutions. Ontological features provides a solution for dealing with the complementary or redundant information related to highlighted genes in a gene set of interest. The *a priori* use of GO subsets allows to highlight the terms in a specific functional category by reducing unnecessary information from enrichment analysis. Nevertheless, defining the functional categories of interest at the beginning of the analysis is not an easy task and requires prior knowledge of the scope of interest. Moreover, these categories often provide GO terms exhibiting a low level of details (*i.e.*, being too generic). Other strategies involve semantic similarities that facilitate the biological interpretation of a gene set that may contain hundreds of genes. Given the significant number of published semantic similarity measures [Pes+09; Guz+12; MCM17], the selection of a metric can be tricky. A recent publication discussed this issue and proposed a classification of these measures according to the type of driven analysis [MM14]. To address these issues, other initiatives using the similarity among the over-represented terms have been proposed. For example, REVIGO [Sup+11] is an *a posteriori* tool that selects only some terms from the output results of enrichment methods. REVIGO reduces the annotation redundancy while it sweeps along the bias of enrichment analysis. Semantic tools presented in section 2.3.2 have the advantage to compute a comparison score

between terms. The extension of such computation to deal with a gene set (that may contain a large number of genes) may be of great interest to synthesize the gene set functional information. However, existing tools are focused on gene-to-gene analysis for grouping the genes sharing similar annotations. Therefore, they do not consider the genes as a gene set in order to provide a global annotation. Thus, they do not take into account the semantic similarity between the annotations of the different genes of a given set. Only the `GOSim` [Frö+07] and `GOSemSim` [Yu+10] tools compute semantic similarity between two GO terms in an automatic way in order to group them according to their similarity using a clustering method. Nevertheless, they only group similar terms without synthesizing the information. The work of Lee et al. presents a new approach for annotating gene sets by using semantic similarity [LHK04]. However, this approach only provides a GO term with a low level of details for each gene set.

Visualization. Despite the diversity and large number of visual metaphors existing to interpret and explore biological questions, this field is still at an early stage [Ker+17]. General challenges related to visualization such as data volume, data type, data representation and interaction still have to be addressed [ODo+10; Mou+18]. Existing visualizations of functional annotation are mainly focused on: (i) adjacency matrices showing the presence/absence of a particular annotation term for a given gene, and (ii) gene-term or term-term networks considering the analysis results or using an external corpus. The first is useful for a direct representation but not for exploring results. The second could be a right approach for exploring data results but node-links can be hard to understand using large hierarchical data due to the crossing edges and the available space that is not optimized [JS91]. Moreover, for large hierarchical data such as trees, planar graphs or DAGs, it is hard to find a visualization algorithm that gives good results in terms of computation time, aesthetic criteria, and emphasized information [GB15]. For that reasons, the use of consolidated systems integrating multiple visual metaphors that show different characteristics of data may be a good alternative. However, the application of this integration is still rare [Ker+17]. In addition, biological visualization generally focuses on a specific task on biological analyses. Moreover, focusing on functional annotation, the current methods supported by visualization of annotation terms of a (narrowed) list of genes are not designed to handle a list of dozens or more gene sets. They are therefore not suitable for a context where the response of the organism to a vaccine or a drug has to be understood as a global mechanism.

Integration of different knowledge resources. Manzoni et al. [Man+16] mentioned key challenges related to data integration. First, the numerous nomenclatures that describe the same thing in different biological databases or resources may result in inconsistencies during the integration. For example, the *BRCA1 DNA repair associated* gene can be associated with multiple identifiers in different databases. Examples of them are:

- HGNC: 1100
- Entrez Gene: 672
- Ensembl: `ENSG00000012048`
- OMIM: 113705

In phenotype knowledge, such inconsistencies are also present. For example, the disease term *multiple personality disorder* is differently described in knowledge resources such as DO, ICD-10, *National Cancer Institute Thesaurus* (NCIt) [Gol+03], *Online Mendelian Inheritance in*

Man (OMIM) [Ham+05], and MeSH. Benites et al. [BSS14] emphasized this challenge caused by the variety of knowledge resources describing the same notion in different ways. For the author, a single resource provides only a certain facet of complex knowledge while integrating several resources may give a more complete information. The second challenge is impacted by the different origins of data. The different sequencing machines and their different processing pipelines provide their results in different formats. This requires additional efforts to make data merging easier. The third challenge relies on the computational performance and the storage capacity due to integration. Because of the large amount of information to be integrated as well as the multi-scale nature and heterogeneity of biological data, ordinary computers are very limited. In consequence, it is necessary to use high-throughput machines, clouds or distributional computational techniques [MPP18]. The fourth challenge expresses the lack of theoretical knowledge in the biological domain and predictor models. For this challenge, Manzoni et al. [Man+16] noticed that there is a lack of models to predict metabolite changes when a pathway is perturbed. Finally, authors mention the absence of efficient pipelines to integrate additional datatypes or metadata to correct biological differences. A last challenge to integrate biological knowledge resources is the lack of relations existing between omics and clinical resources. In particular, for a given gene or gene set, there is no resource that provides the inter-relations between biological processes or pathways with diseases or clinical phenotypes.

Throughout this manuscript, we propose alternatives or solutions in order to address some of the challenges presented in this chapter.

Chapter 3

Gene sets visualization

As presented in section 2.3.3, managing the large number of annotation terms associated with a gene set is usually difficult. To address this issue, statistical methods, called enrichment methods, have been proposed [HSL09; KSB12; Thé+15]. These tools show an important pitfall related to the redundancy among the results [Sup+11], resulting from the non-use or the under-utilization of semantic relations between annotation terms. Structures of knowledge resources such as GO may be used to increase the biological information while reducing the redundancy by taking into account the relations between terms. Whereas the hierarchical structure has been exploited by methods such as `GOSim` [Frö+07], `G-SESAME` [Du+09], `GOSemSim` [Yu+10], `GFSAT` [Xu+13] and `GOGO` [ZW18] that aim to compare two genes or two terms, only few others address the gene set annotation problem [Bau+08; Sup+11; Thé+15].

Although potentially useful, the use of the hierarchical structure between annotation terms increases the amount of information to be dealt with. The size and complexity of these data urge the need of dedicated visualization techniques in order to interpret such information. However, the choice of the adequate visual metaphor is an arduous task. Consequently, the addition of interactive solutions may help to understand different layers of analysis (*e.g.*, a global view that represents the main biological processes in which many genes are involved versus a specific view that shows up few genes with detailed information). But, applying an inappropriate visualization may confuse users and may result in wrong interpretations. For example, a visualization of a set of data with similar values within a pie chart could be hard to interpret. Conversely, using other kinds of visualizations, such as a barplot, could be more adequate than a pie chart. As commented in section 2.3.3, focusing on the functional annotation of a gene set, two visual metaphors have been widely used [Mou+18]: node-link diagrams and heatmaps. However, these metaphors may not be the most pertinent. Heatmaps can be very useful when large data are explored, but because it is a visualization in the form of a matrix, it only allows the visualization of two facets of data: the one placed in rows and the other in columns. Node-links diagrams are adapted to represent hierarchical data but, due to their low space utilization, they can be inefficient if the data to be explored are large [JS91]. Moreover, if the visualization of a gene set can be hard, the complexity is increased when dozens of gene sets are studied.

Most bioinformatics enrichment tools mainly annotate a given gene set and the visualizations proposed by these tools are very limited for dealing with multiple gene sets. Some authors developed techniques in order to create gene set repertoires to describe their impact on diseases [CB14], their interactions with vaccine trials [Li+13] or simply the main genes that over-express into immunological cells [Bin+13]. For annotating multiple gene sets, each

gene set must be annotated one-by-one, thus generating a higher amount of information to be processed. To the best of our knowledge, there is no visual metaphor that represents the functional annotation results of multiple gene sets.

This chapter is addressing the main question: "*what differs in terms of visualization when exploring the annotations from one to multiple gene sets?*" by taking into account an other important issue: "*how the knowledge resources' structure may help for the visual interpretation?*" In order to investigate these questions, three visualization prototypes have been developed.

- The two first prototypes aim to represent the functional annotation of a single gene set without considering the DAG structure. These visual metaphors are the results of the developments made by master students whom I supervised. This exploratory work had the double benefit of investigating visual metaphors within the classical framework of gene set annotation (using enrichment methods) and of experimenting the supervision of students.
- The third visualization tool is the result of an international collaboration with colleagues of the University of Murcia in order to reconcile different knowledge resources used for annotating multiple gene sets by using lexical similarity and has been presented during an international conference in *Information Visualisation* (IV). This visualization prototype presents a combination of a space-filling visualization and an indented tree involving the GO structure for displaying and exploring the results.

Before visualizing the annotation results, annotation from one or multiple gene sets has obviously to be computed. In section 3.1, we present a simple framework that has been developed based on enrichment analysis to carry out the two visualization prototypes. Section 3.2 involves the third visualization tool in which the use of enrichment analysis methods is still central in the framework but capitalize on lexical approaches to address the scale-up resulting from the visualization of the annotation of multiple gene sets. The latter approach is in consequence more detailed. We conclude the chapter in section 3.3.

3.1 Visualizing the annotation of a single gene set

With the objective to visualize the annotation of a single gene set, a pipeline has been developed. The main objective in this preliminary work for the development of `GSAn` (see chapter 4) was to investigate the different visual metaphors that may be relevant in the context of gene set annotation. For this reason, most of the following developments have been based on enrichment analysis methods to provide gene set annotations. The proposed pipeline is composed of four steps (figure 3.1). The first step uses GO terms to annotate a gene set of interest using the enrichment analysis carried out by `g:Profiler` [Rei+07] without post-treatment. Then, the comparison of the GO terms are computed and stored in a semantic similarity matrix. Thirdly, a clustering stage is applied in order to generate a partition of "similar" GO terms. Finally, for each group of GO terms, a simple process selects the *Most Informative Common Ancestor* (MICA) [Res95] term of the group by making use of the GO structure. At last, two visual metaphors have been proposed in order to facilitate the interpretation of results.

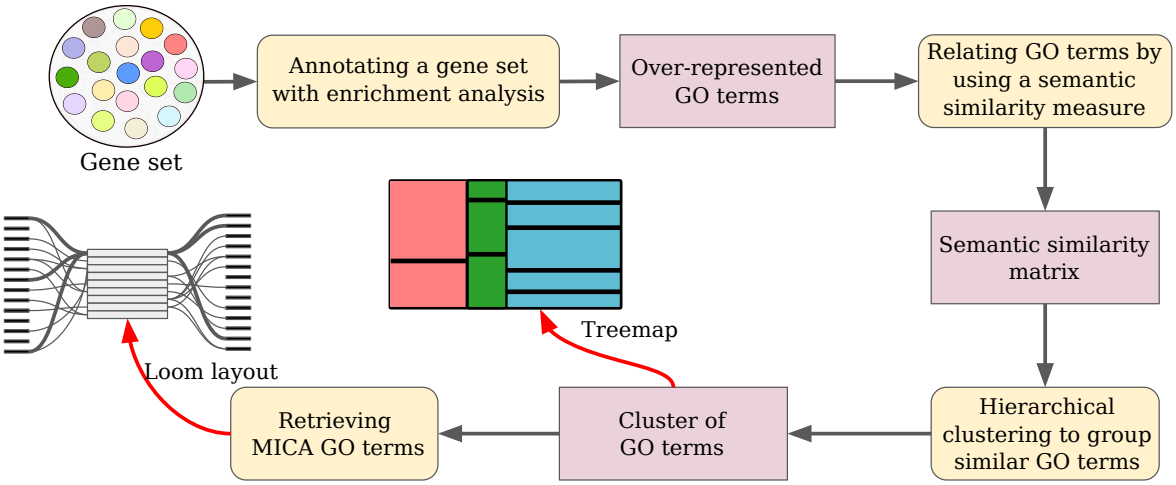


FIGURE 3.1: Implemented pipeline to annotate and visualize a single gene set. The pipeline is composed of four steps (colored in yellow) generating a new transformed data (colored in pink). From the last two steps, two visual metaphors have been proposed (red edges): a loom layout based on the *Most Informative Common Ancestor* (MICA) terms and a treemap displaying clusters of GO terms.

3.1.1 Computation of the gene set annotation

Annotating a gene set by using enrichment analysis methods. The first step involves the annotation of gene sets by using the enrichment analysis. Thanks to these methods (described in section 2.3.1), lists of over-represented annotation terms were recovered according to their statistical use trough the genes within the sets. As this research work focuses on the visualization method, we arbitrarily chose `g:Profiler` [Rei+07; Rei+16] (categorized in the Single Enrichment Analysis or SEA class in section 2.3.1) for recovering annotations. We only selected annotations corresponding to the sub-ontology *Biological Process* (BP).

Relating GO terms by using a semantic similarity measure. As mentioned in section 2.3.2, semantic similarity measures are relevant to group annotation terms into categories according to their significance and then to guide the reduction of their number. Thus, some tools were developed for this purpose [Frö+07; Wan+07; Yu+10]. Then, to compute *a posteriori* the semantic similarity between pairs of each of the over-represented terms, we used `GOSemSim` [Yu+10] with default settings. Five types of semantic similarity measures are implemented in `GOSemSim`, among which four of them use an *IC* score based on the annotation corpus [Res95; JC97; Lin98; SA07] and one measure involves the *semantic value* introduced by Wang et al. [Wan+07]. Again, studying the impact of using a similarity measure rather than another was not the objective of the present work and will be deeply investigated in the chapter 4.

Hierarchical clustering of similar GO terms. In order to group terms according to their semantic similarity, hierarchical clustering was applied. The choice of an unsupervised hierarchical clustering was motivated by the presence of hierarchical relations within GO whose

structure is a DAG [SSZ04]. The optimal number of clusters was determined based on the computation of the *Average Silhouette Width* (*ASW*) score [BGL11], which is described in section 4.2.3.

Retrieving the MICA of a group of GO terms. At last, a step was performed to reduce the number of terms and to calculate the best candidate to represent each cluster of terms. To do so, a MICA term was assigned to each cluster by using `GOSim` [Frö+07].

3.1.2 Visual metaphors to represent the annotation of a single gene set

Two visualization developments have been achieved by students under my supervision to explore the output of the previous pipeline.

- The first visualization was implemented by four master students in bioinformatics. They developed a space-filling layout in the form of a treemap in order to represent the results of the clusters of terms.
- The second visualization was implemented by a first year student of the *ENSEIRB-MATMECA* engineering school. She developed a node-link diagram by applying a loom layout[1].

Drawing a treemap

This visualization was developed to represent a cluster of similar GO terms. Inspired by Johnson and Shneiderman [JS91], this visual metaphor (developed with the `FaTuM` library [PA15]) is a treemap drawn according to the *slice-and-dice* algorithm. The treemap visual metaphor allows to distribute the different entities that are hierarchically related to each other into a limited space. The complete hierarchy of GO was not considered in this work. Instead, a hierarchy of two levels was used considering the clusters as the roots and the terms involved in the clusters as their descendants. By using a given space, rectangles with a variable size and a variable color represent the attributes of each element (*i.e.*, cluster or terms). More precisely, the color corresponds to the cluster while the size is related to the number of genes that are associated with a GO term.

The slice-and-dice algorithm draws a first rectangle that is split into smaller rectangles according to the cluster of similar terms obtained from the pipeline framework. Then, each sub-rectangle corresponding to a cluster is in turn sub-divided according to the number of GO terms within the cluster. Figure 3.2A illustrates the partitioning of a treemap that represents annotation clusters. Thus, the initial rectangle represents the totality of the data, a first vertical division of this rectangle highlights the clusters, and then, horizontal divisions display the terms of each cluster. As can be observed in this figure, we can see that *Cluster 1* has an area that corresponds to five over fifteen genes in total and that *Term_c1 1* from this cluster annotates three out of five genes.

To support the exploration within the treemap, our prototype integrates an expanding interaction tool that provides a *focus-plus-context* functionality to deal with the following two

[1]The concept *loom layout* has been defined by Nadia Bremer in her blog `VisualCinamon` (https://www.visualcinnamon.com/2017/08/d3-loom)

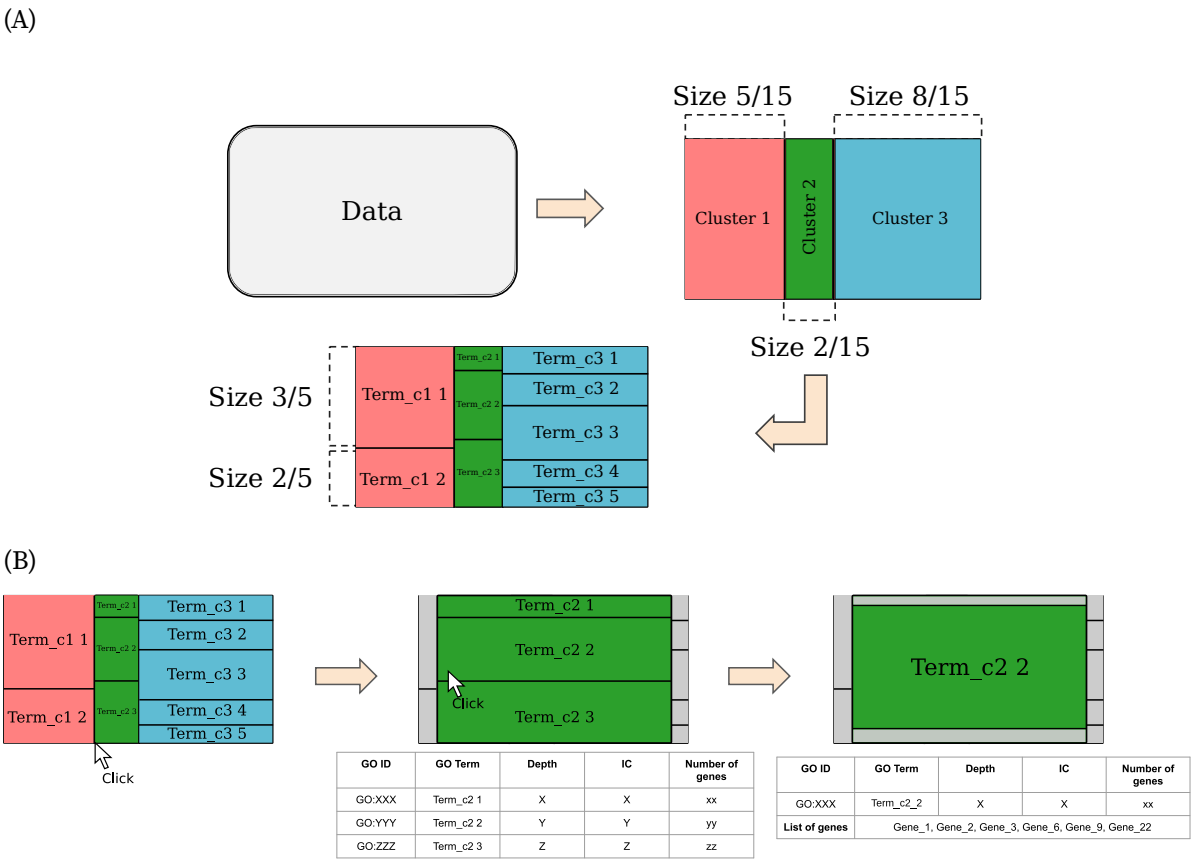


FIGURE 3.2: (A) A schema representing the creation process of the treemap according to Johnson and Shneiderman [JS91], by using the gene occurrence to compute the size of the rectangles in the treemap. (B) The interactions proposed in the visualization in order to display a cluster or a GO term.

levels (figure 3.2B): clusters and terms. By clicking on a cluster, the rectangle corresponding to the cluster is expanded, thus showing the terms involved within the cluster. The color of the expanded cluster and its terms is maintained while the rest of the cluster and terms are shrinked and turned into gray in order to highlight the focused cluster. Additionally, a table describing the information of the involved terms is provided below the treemap. This table represents the characteristics of terms, *i.e.*, their depth, their *IC*, but also the gene occurrence within the gene set. When clicking on a GO term, the same actions (*i.e.*, expansion and color) are produced. In contrast to the cluster expansion, the generated table below the treemap contains only the list of genes in the gene set associated with this GO term.

Drawing a loom layout

The last step of the pipeline assigns a MICA term to each cluster of terms. Unlike the previous visualization, three sets of elements can be visualized within the proposed visual metaphor: (i) the genes from the gene set that are annotated by at least one GO term, (ii) GO terms that annotate these genes, and (iii) the MICA term of each cluster of similar GO terms. They are represented together using a loom layout. The advantage of the prototype visualization is that it allows both to show an overview summarizing the pipeline results and to explore the pipeline results in details thanks to interactions. The overview displays the gene set annotation by representing the MICA terms as rectangles in the center (called *internal entities*) and

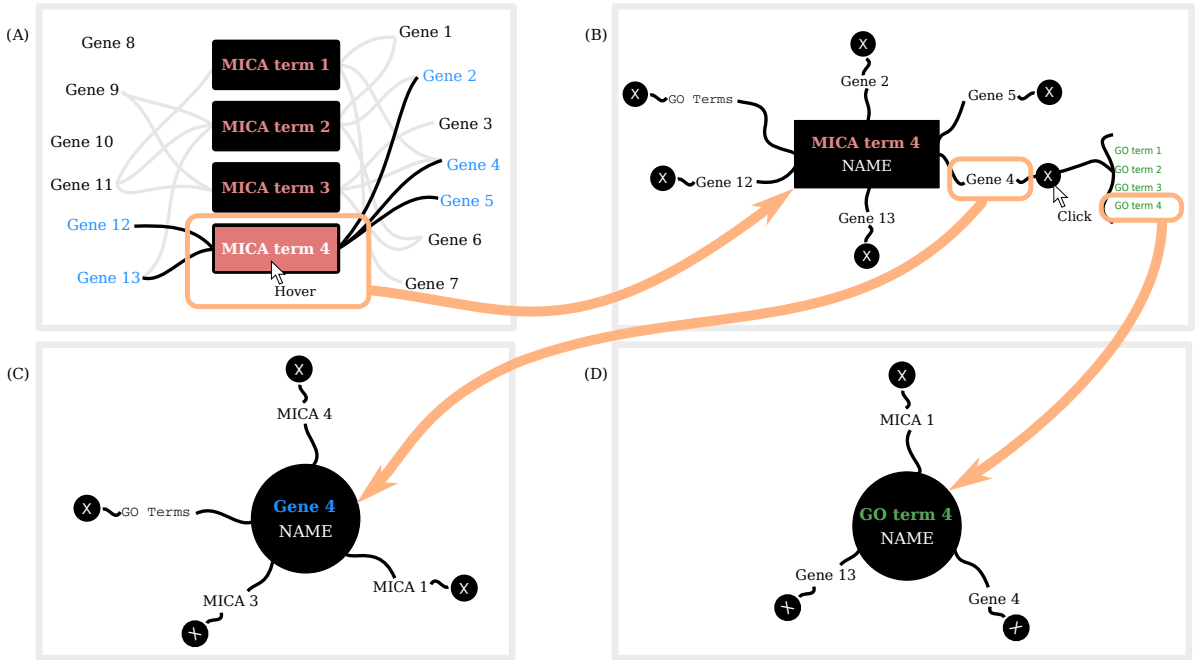


FIGURE 3.3: Visual description of the loom layout according to two levels of representation: (A) A general representation shows the overview of results where the MICA terms are the internal entities and their related genes are the external entities. Hovering over a MICA term, the corresponding rectangle is colored in salmon, the associated genes are colored in blue and the edges that are not connected with the chosen MICA term are colored in light gray. Three possible detailed views are available for the three types of elements: (B) a MICA term, (C) a gene, and (D) a GO term. A detailed representation of any of these types has a single internal entity and shows its connections. Additionally, clicking on the black circle associated with an external entity allows to see the list of GO terms associated with both this external entity and the internal entity. The orange arrows show an example of click interactions that enable to navigate within the different types of views.

their links to the genes of the gene set, called *external entities* (figure 3.3A). When the mouse pointer is on a MICA term (*hover* action), the rectangle is colored in salmon, the associated genes are colored in blue and the edges that are not connected to the MICA term are dimmed in light gray. A detailed view can be obtained by interaction facilities for any type of element: a MICA term, a GO term or a gene. For the detailed views, the loom layout metaphors of the overview have been declined by adapting the internal and external entities. All detailed views present a single internal entity corresponding to the element under investigation, as well as clickable elements (black circles) that are associated with external entities for showing the list of GO terms associated with both the external and the internal entities. Moreover, one can navigate between these detailed views through interaction actions. The different detailed views illustrated on figure 3.3 are described hereinafter.

- Panel B: the first interaction tool consists in clicking on a MICA term (`MICA term 4` in figure 3.3A) from the overview to get a detailed level of information. The resulting view shows the MICA term as the internal entity placed at the center with its related genes and the involved GO terms (included into the entity labeled `GO Terms`) placed around. By clicking on a black circle, one can get the list of GO annotation terms from the cluster annotating the gene of interest (`Gene 4` in figure 3.3B).

- Panel C: this view is obtained by clicking on a gene. A new loom layout is represented where the internal entity is the gene and external entities are the MICA terms and the

involved GO terms that annotate this gene (figure 3.3C).

- Panel D: this view is obtained by clicking on a GO term. A new loom layout is represented where the internal entity is the GO term and external entities are the genes that are annotated by this GO term and its associated MICA term (figure 3.3D). In this view, the black circle of an external entity represents *all* its associated GO terms.

The three types of detailed view (MICA term, gene, and GO term) can be accessed from any other view. Moreover, from any view, a click on any internal entity opens the corresponding Web page of the *National Center for Biotechnology Information* (NCBI[2])if the entity is a gene, or of `QuickGO`[3] if the entity is a MICA or GO term.

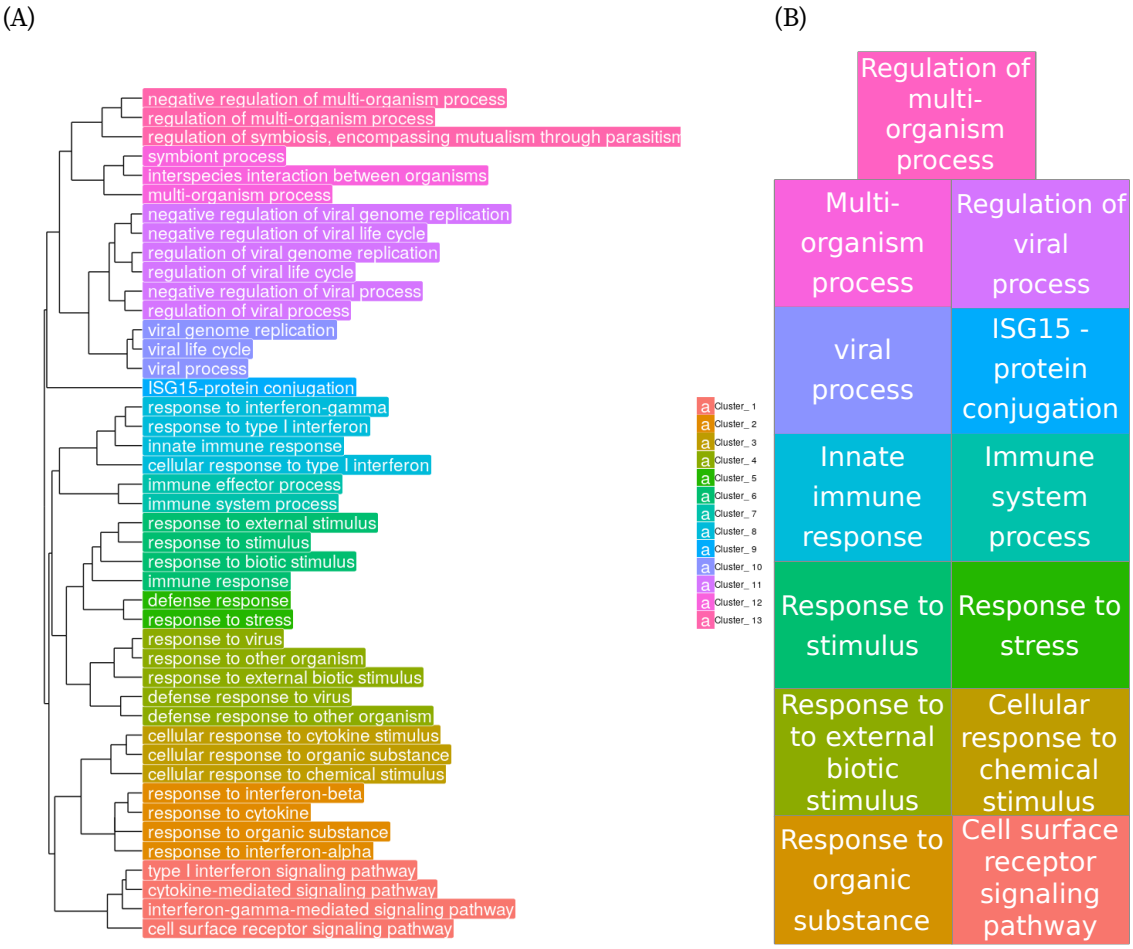


FIGURE 3.4: Results of the annotation for the gene set annotated as *Interferon* by Chaussabel and Baldwin [CB14]. (A) The clustering step results in thirteen clusters represented in the dendrogram. (B) The last step of the pipeline provides the MICA term of each cluster.

3.1.3 Case study

To illustrate the effectiveness of visualization prototypes in exploring the annotation of a single gene set, we performed an analysis making use of a gene set from the dataset [`C-260`] that

[2] `https://www.ncbi.nlm.nih.gov/gene`
[3] `https://www.ebi.ac.uk/QuickGO/`

involves 27 genes annotated as *interferon* by Chaussabel and Baldwin [CB14] (that we already used for providing visual examples in figure 2.6). Overall, 115 GO terms involved in BP are associated with at least one of the genes in the set according to GOA. By applying the analysis pipeline, the enrichment step produced 44 over-represented GO terms while the clustering resulted in thirteen MICA terms (one GO term for each cluster).

Figure 3.4A shows a dendrogram created after the hierarchical clustering of the similarity matrix and figure 3.4B displays the MICA term of each cluster. These thirteen MICA terms actually present redundancy. Indeed, *innate immune response* is a descendant term of *immune system process* and *response to stimulus*, the latter being the ancestor of every GO term containing the word *response*.

Figure 3.5A shows the GO terms part of *Cluster_10* that corresponds to viral processes. Thus, *viral life cycle* (`GO:0019058`) and *viral genome replication* (`GO:0019079`) are the terms having the highest *IC* and number of associated genes.

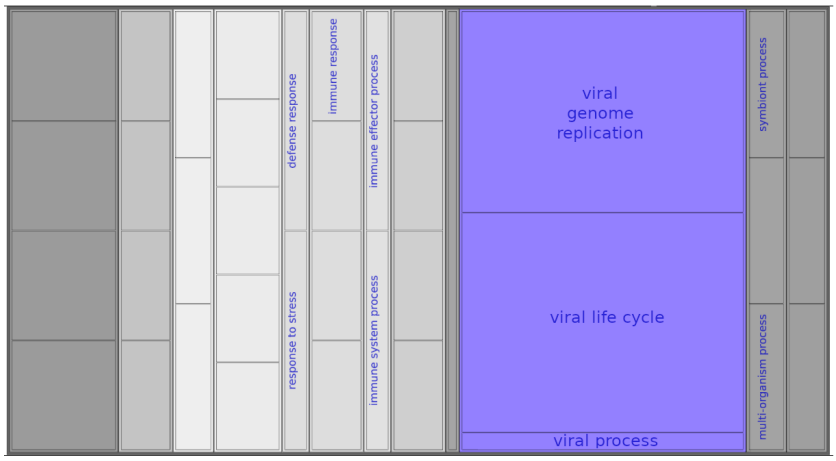
The treemap visualization only uses color as main attribute. Moreover, due to the fact that the space to draw the treemap is limited, some labels are hidden in the overview. This presents visual limitations, in particular for large clusters for which many information have to be displayed. In contrast, the loom layout provides a put-on-context by means of the MICA terms (figure 3.5B). Thus, in the middle of the global visualization, thirteen GO terms (MICAs) represent the clusters which involve processes like *viral process*, *cell surface receptor signaling pathway*, and *innate immune response*. When hovering over a MICA term or a gene, the related genes or MICA terms are shown, respectively. When clicking on one of these entities, the visualization displays the details of the clicked entity. For example, figure 3.5C shows the details of the `IFITM3` gene. This gene is involved in twelve out of the thirteen MICA terms. The small black circles show the number of GO terms associated with a MICA term and that annotate the chosen gene. By clicking on one of these circles, the corresponding GO terms are displayed (*viral process* in figure 3.5C).

Despite the global view offered by this visualization that resumes the gene set and enables to explore the results, it presents two main limitations. First, there is a loss of context when clicking and focusing on a given entity and when repeating again and again this process if users want to explore all the entities. The second limitation concerns missing information. First, the visualization shows the relations between GO terms and genes but does not provide information about attributes, such as the IC or the depth of a GO term. On the other hand, when exploring the MICA terms, it appears necessary to show their position within the GO structure so that users may know how a GO term is related with a given MICA term.

3.2 Visualizing the annotation of multiple gene sets

To the best of our knowledge, there is no available bioinformatics tool enabling the visualization of the functional annotation of multiple gene sets (section 2.4). It is noteworthy that the database gene set repository `GeneWeaver` provides analysis tools that visualize multiple gene

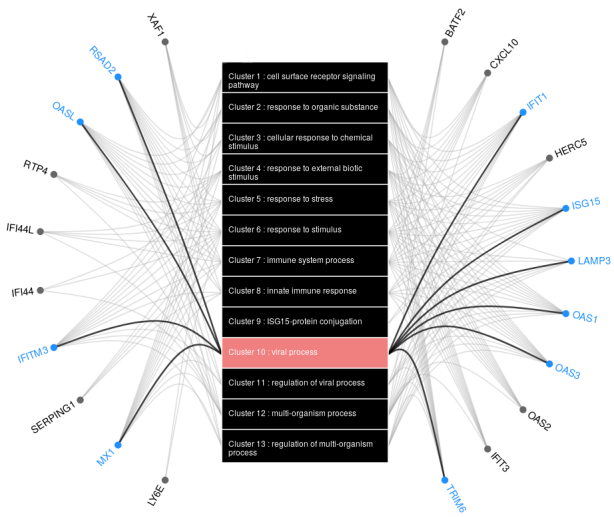
(A)



CLUSTER_10

Term Name	GO term	IC	Depth	Gene number
viral process	GO:0016032	2.17	4	2
viral life cycle	GO:0019058	3.78	5	21
viral genome replication	GO:0019079	3.78	5	18

(B)



(C)

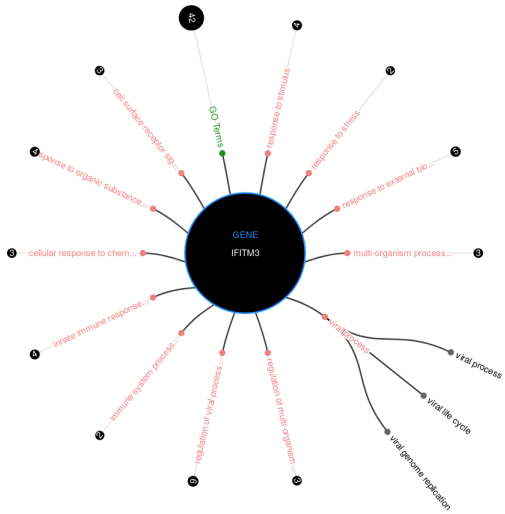


FIGURE 3.5: Proposed visualizations of the annotation of a gene set. (A) The presented treemap is focused on *Cluster_10* and the table below shows the details of the involved GO terms. (B) The loom layout provides an overview of MICA terms and the genes they annotate. (C) The loom layout presents the details of the `IFITM3` gene (the numbered black circle corresponding to the external entity *viral process* has been developed).

sets [Bak+11]. However, even if these tools aim to provide a visualization that shows the closeness of gene sets sharing similar genes, the functional annotation is not taken into account.

In section 3.1, a single gene set was considered while the information from the GO structure was discarded. The present section aims to propose a pipeline to explore multiple gene sets

according to the GO structure. The visualization of the DAG structure of GO can be realized by an indented tree (by duplicating the nodes) or by a node-link diagram. Both visualizations have been compared in usability studies of ontologies [FNS13; FNS17]. By comparing several factors (effectiveness, efficiency, workload, usability, and qualitative feedback), it has been reported that indented trees are more readable for novice users while node-links diagrams are more intuitive and avoid visual redundancy. Also, indented trees are more efficient in searching information and node-link diagrams in processing information. Graham and Kennedy [GK07] consider node-link diagrams and space-filling as basic layout styles to visualize a DAG or a tree structure. However, as mentioned at the beginning of this chapter, node-link diagrams can be hard to understand when using large hierarchical because edges may cross each other and the space is not used in an optimal way [JS91]. Moreover, the hierarchy shown in treemaps is arduous to detect. For these reasons, finding an algorithm that provides a good representation of the hierarchical data is a difficult task.

This research work has been led in collaboration with Jesualdo Tomás Fernández-Breis and Manuel Quesada-Martínez from the University of Murcia and Romain Bourqui from the University of Bordeaux. The aim of this work was to make use of multiple knowledge resources in order to improve our understanding of life. The manual expertise to clearly decipher the main functions that may be related to gene sets is time-consuming and becomes impracticable when the number of sets increases, as it is the case in vaccine/drug trials. In this frame, a pipeline has been developed in order to annotate gene sets according to different knowledge resources and to map these resources to GO, thus homogenizing the annotation.

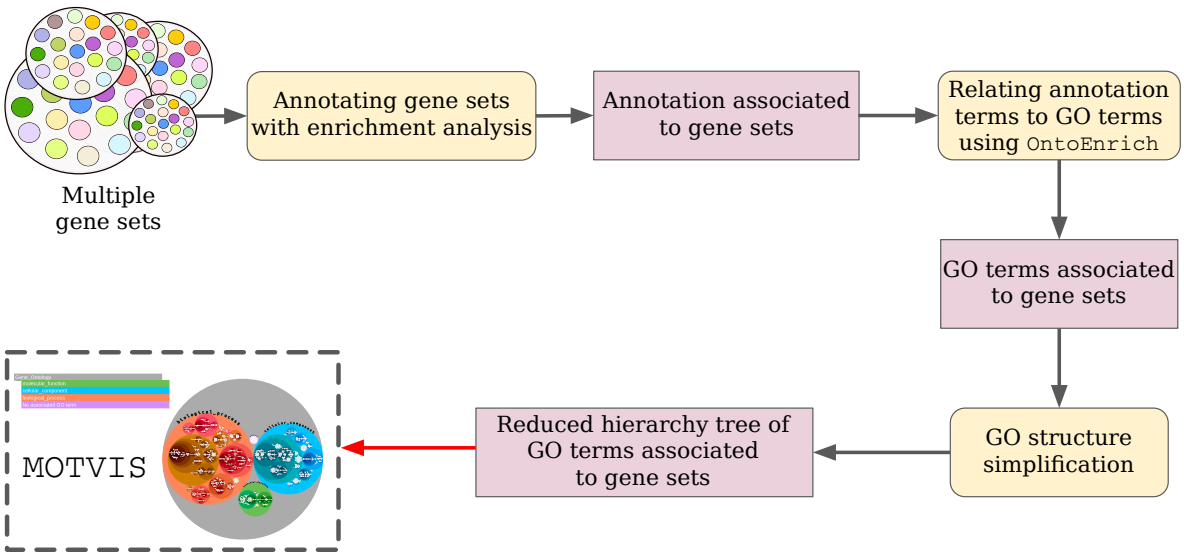


FIGURE 3.6: Pipeline to annotate and visualize multiple gene sets. The pipeline combines enrichment and annotation simplification for connecting gene sets and selected GO terms. The pink rectangles represent input or output data for a given method in a given step (represented by the yellow rounded rectangles). For the last step, the visualization prototype MOTVIS (*MOdular Term Visualization*) has been proposed (red edge).

The proposed workflow consists in three main steps (figure 3.6). The first step aims at annotating multiple gene sets using a limited number of annotation terms. Then, these terms are lexically processed in order to identify close terms within the GO. Next, as the GO is large, a simplification step is performed to select only terms which are relevant for the interpretation of the input gene sets. Finally, the last visualization prototype was used in order to represent the results by using the simplified GO structure.

3.2.1 Annotating gene sets

In this work, we also used `g:Profiler` [Rei+16; Rau+19] as it makes use of several annotation databases. Using different knowledge resources allows to obtain the functional roles of the gene sets from different points of view (as gene annotations may have been done at different cell organization levels). However, such a tool provides a given gene set with annotations coming from various resources (*e.g.*, GO, KEGG, REACTOME, WikiPathways, TRANSFAC, and HPO) even if these annotations describe the same (or similar) biological functions. This drawback is handled by the next step of our pipeline, which reconciles the output annotation terms with terms from GO, used as a reference ontology. Note that this step also makes it possible to use any enrichment tool that provides an annotation from multiple resources.

3.2.2 Relating annotation terms to GO terms

For reconciling the annotation terms provided by the g:Profiler with the GO terms, a lexical approach implemented within the `OntoEnrich` framework was applied [Que+15a]. Annotations were processed and `OntoEnrich` was configured for: 1) decomposing annotations in tokens based on the tokenization and lemmatization strategies proposed by the `Stanford` *Natural Language Processing* (`NLP`) `toolkit` [Man+14], 2) searching groups of consecutive annotation tokens that correspond to the whole label of a class in the ontology or any of its synonyms (considering related, narrow and exact synonyms in GO), and 3) defining and applying two filtering strategies. The first filter removes the most general terms if any of its descendants are also associated to the matched annotation. The second filter is based on the lexical relations between terms. In practice, we removed the mappings involving terms that are contained in others. In summary, the inputs of this step are textual descriptions of annotations that are automatically converted into semantic annotations, being GO terms. As an illustration, if a given gene set is annotated by the annotation term *cell cycle, ATP bindings*, `OntoEnrich` identifies four partial matches in the GO: *cell*, *cell cycle*, *ATP binding* and *binding* but only *cell cycle* and *ATP binding* are kept.

3.2.3 GO structure simplification

Once the list of GO terms was obtained using `OntoEnrich`, a filtering stage was performed to generate a subgraph with only the most pertinent GO terms. To do that, for each GO term obtained by `OntoEnrich`, their most informative parent term was recovered and this process was recursively applied until the root term was reached (figure 3.7). For determining the most informative parent, the following IC measure was computed for each selected GO term t [ZWG08a]:

$$IC_{zhou}(t) = k \cdot (1 - \frac{log(freq(t))}{log(freq(root))}) + (1-k) \cdot \frac{log(d(t))}{log(MDO)} \tag{3.1}$$

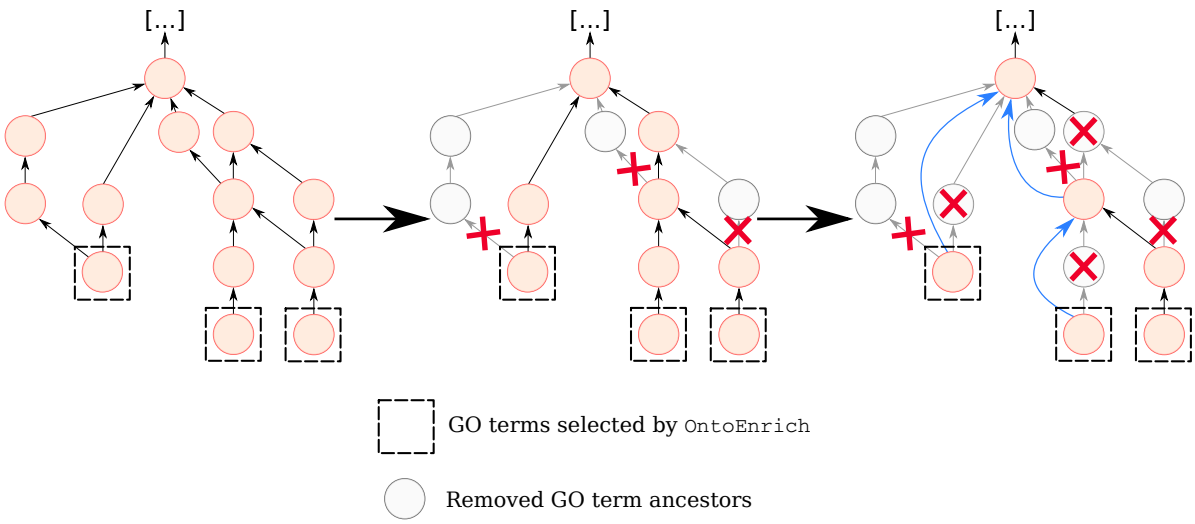


FIGURE 3.7: Illustration of the process of GO simplification. This example shows at the left an extract of the GO DAG showing three terms provided by `OntoEnrich` and some of their ancestors. During the first step (from the left to the middle), if a GO term has more than one parent, the relation with the less informative parent is discarded (represented as a red cross on the edge connecting the GO term and its parent term). Therefore, if this parent is not associated with any GO term provided by `OntoEnrich`, it is removed and so are all its ancestors. Then, from the middle to the right, a second step is processed to remove the ancestors that are connected to only one GO term provided by `OntoEnrich` as well as general ancestors that cover the same GO terms as another more specific ancestor. A new link (colored in blue) represents the connection between two related GO terms if their intermediate GO terms were removed.

where $freq(t)$ is the number of descendant terms of t and $d(t)$ is the depth of t within the ontology. MDO is the maximal depth in the ontology and k is an adjustable factor providing a weight for each item of the equation. This factor can be adjusted so that the equation provides more priority to descendants (when k is near 1) or to depth (k near 0). In our experiments, we chose 0.5 as proposed in Zhou et al. [ZWG08a]. After having generated this subgraph, to avoid visual issues that complicate the exploration of results (*i.e.*, many rings may cover a single GO term), as shown in figure 3.7, only the GO terms being ancestors of at least two GO terms obtained by `OntoEnrich` were kept. Moreover, if two hierarchically related ancestors cover the same GO terms, the most general was removed. This way, the DAG is transformed into a tree, making it to be represented in the visualization prototype described in the next section.

3.2.4 Visualizing with `MOTVIS`

The output of the analysis pipeline described in the previous section is a hierarchical tree of GO terms whose leaves are associated with the input gene sets. Considering that each gene set is linked to its GO terms in that tree (therefore forming a new DAG), we decided to duplicate each gene set to unfold that DAG into a tree whose leaves are gene sets. Such an idea has already been used by Koenig et al. [Koe+07] and Tsiaras et al. [TTT09] to support the visualization of DAGs. If two GO terms which are hierarchically-related in the tree annotate the same gene set, the association with the most general GO term is removed. This technique offers the advantage of simplifying the representation while the duplication of leaves (*i.e.*, gene sets in our case) may lead to misinterpretations. In the following section, we present how the

produced tree is visually represented, how we support the identification of duplicated gene sets and the interaction tools supported by our prototype called MOTVIS (*MOdular Term Visualization*).

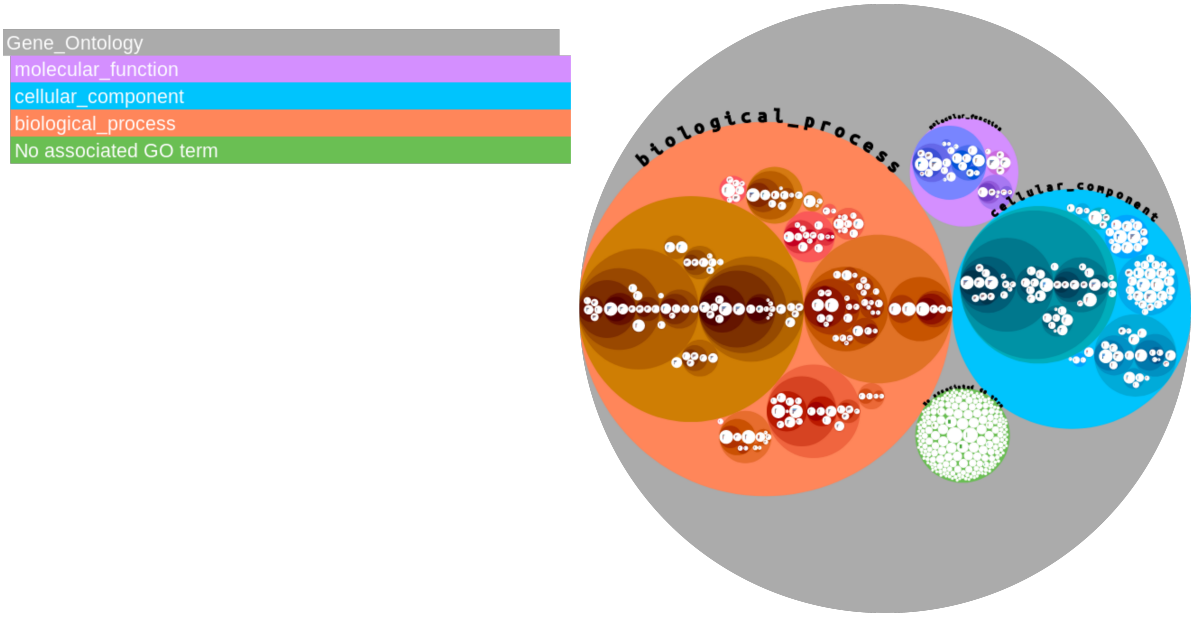


FIGURE 3.8: Indented tree and treemap view representing the immune response within the [C-260] dataset. The four highest level nodes correspond to the 3 ontologies of GO: *Biological Process* or BP (colored in orange), *Cellular Component* or CC (in blue), *Molecular Function* or MF (in purple), together with all gene sets which are not annotated by any GO term (in green).

Drawing the association between gene sets and GO terms

Inspired from [Koe+07], MOTVIS uses two interactive and interconnected views (developed with the D3 library [BOH11]): a circular treemap and an indented tree. Unlike Koenig et al. [Koe+07] and Tsiaras et al. [TTT09], a circular treemap was preferred over rectangular or Voronoi treemaps. Indeed, circular treemaps present the following advantages in comparison with other kinds of treemaps [Wan+06; ZL15]: (i) there is a clear separation between nodes, and (ii) the different levels of hierarchical data are easily interpretable. Nevertheless, this is made at the expense of a larger space usage. Thanks to our annotation strategy, the generated tree contains few hundred nodes which counterbalances that drawback.

For the annotation of gene sets, the number of genes in each set is an important information as a large gene set (*i.e.*, containing a large number of genes) may have a more generic biological role in the cell overall functioning. Therefore, larger gene sets are emphasized by setting their size proportionally to the number of genes they contain. A collapsible intended tree view has also been developed to display additional information to what is displayed within the circular treemap. In that view, each node (GO term and gene set) is represented as a rectangle and clicking on one of these rectangles expands it and displays its child nodes. This helps users when seeking for a GO term of interest and thus for a particular functional role that could have been impacted during the experiment (see figure 3.8 to observe an overview of the three sub-ontologies BP, MF and or CC). Such a visualization technique was preferred

over other techniques like node-link diagrams (as in [Koe+07]) because it has been shown that it improves users' experience in usability studies of ontologies [FNS13; FNS17].

Overcoming the duplication issue

As mentioned above, the gene sets were duplicated and linked to each of their annotating GO terms. Such duplication may hinder a good understanding of the biological roles of gene sets. For facilitating the identification of duplicated gene sets, we combined a specific tree node coloring algorithm and bar charts within gene set nodes.

First, tree nodes were colored using the `TreeColors` algorithm [TJ14] that assigns similar colors to close nodes. The basics of this algorithm are to use the HCL (*Hue-Chroma-Luminance*) color space and to recursively divide a hue interval associating a node with its children (the hue interval of the tree root is set to $[0, 359]$). Then, increasing the chroma and reducing the luminance according to the depth of a node improve the perception of depth within the tree. This algorithm was applied to the entire tree, except for the leaves that represent the gene set nodes. Setting their color to an unused color facilitates the identification of gene set nodes in the tree. Moreover, within each gene set node, a bar chart is displayed representing its annotation terms (using their assigned colors) and the level of confidence according to the *p-value* computed by the enrichment analysis. It allows to identify gene sets annotated by several GO terms, as well as to provide the number of times a gene set is duplicated and the level of its annotation terms in the hierarchy using colors.

This is an important feature of our visualization because it emphasizes gene sets annotated with several similar terms that could improve the level of confidence of users. Nevertheless, an important drawback is presented when considering large trees because some of the node colors may be perceptually similar (as mentioned in [TJ14]).

Interaction facilities

To support the exploration in the proposed visualizations, `MOTVIS` integrates several interaction tools. Besides a classical *zoom* interaction tool that supports basic exploration, it also integrates an interaction that allows to focus on a GO term or a gene set of interest. Clicking on a node (either a GO term or a gene set), in the treemap view, automatically zooms on that particular node of the tree. As mentioned in section 3.2.4, the two views being interconnected, it also expands the indented tree to display the entire path between the root node and the focused one (and the remaining expanded paths of the tree are collapsed). If the clicked node represents a gene set, all paths between the root node and all duplicates are expanded. Clicking on a node in the indented tree view also allows to zoom on the corresponding duplicate gene set in the treemap view. A last interaction tool thus tries to alleviate this drawback and highlights all duplicates in the treemap view when hovering over a gene set annotated by several GO terms.

3.2.5 Case study

To illustrate the usefulness of our pipeline and `MOTVIS`, we carried and experimented two analyses making use of the [`C-260`] dataset related to the immune response within the human

population. The objective of such analysis is to identify the key regulators that manage the immune defense system.

FIGURE 3.9: Zoom on the gene sets whose annotation provided by the enrichment analysis has been completely or partially mapped to the GO term *signaling* when `OntoEnrich` was applied. All the modules (Mx.x) or gene sets that are annotated by this GO term are presented as white circles within the pink circle that corresponds to the *signaling* annotation term. Moreover, the histogram within each white circle gives additional information for each gene set (Mx.x).

Applying our analysis pipeline, the enrichment steps produced 1,296 annotation terms while the alignment with the GO and simplification steps allowed to reduce drastically their number to 157. figure 3.8 shows the treemap view representing these gene sets together with the simplified GO structure which corresponds to a tree of 782 nodes (including all duplicates of gene sets). In immunology, one of the most important biological activities comes from the interaction and communication between cells. Both processes can be investigated by studying the signaling pathway from a global point of view (which corresponds to the signal transduction by which a signal is transmitted from cell to cell). The first step of our exploration was therefore to search and select the *signaling* annotation term from the indented tree view. The resulting view displays seven gene sets (figure 3.9). The indented tree shows the annotation terms computed by the enrichment tool and our reconciliation step has related them to the *signaling* GO term, which seems relevant here. Looking at the bar charts within the gene set (in the treemap view), one can identify some gene sets where the level of confidence of the *signaling* annotation term is high (*e.g.*, M1.2 and M3.4) because the corresponding bar is large, while others are low (*e.g.*, M5.1 and M5.9). This is confirmed by the study of Chaussabel and Baldwin [CB14] who manually annotated M1.2 and M3.4 as *interferon* (which is closely related to *signaling*) while M5.1 and M5.9 were annotated by *inflammation* and *protein synthesis*, respectively.

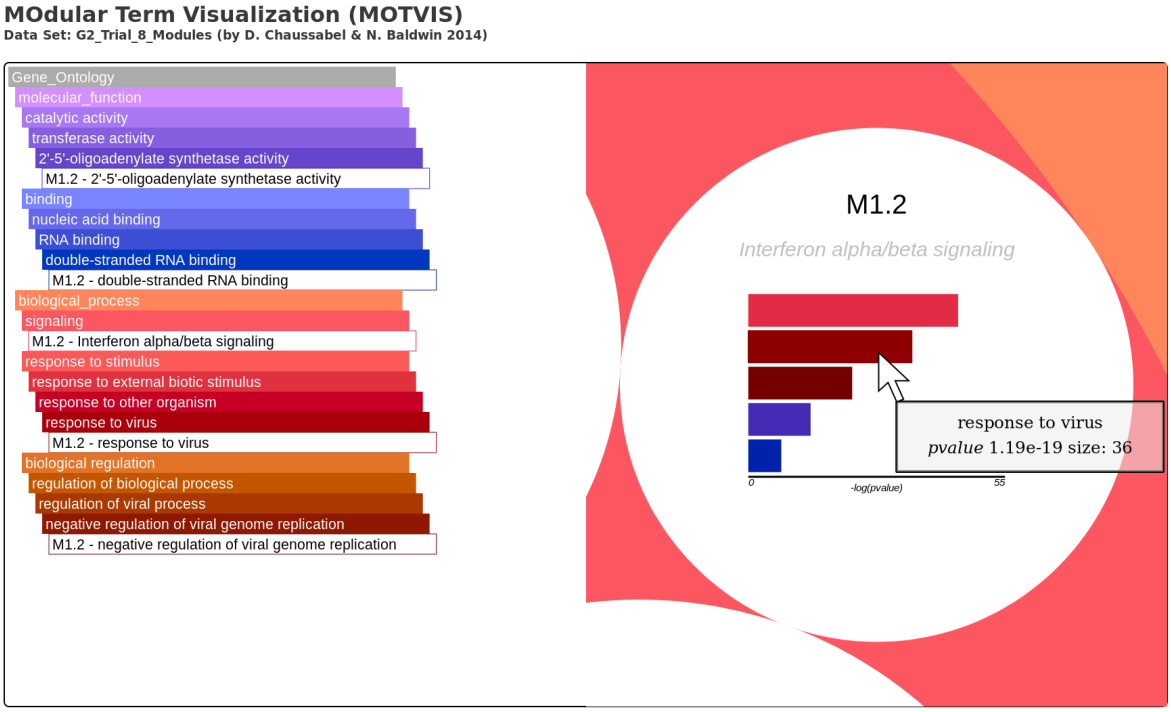


FIGURE 3.10: MOTVIS view focused on the M1.2 gene set. The histogram displays bars corresponding to the other terms (in their appropriate color) annotating this gene set, the width of each bar depending on the *p-value* obtained from the enrichment analysis. Hovering the mouse over each bar, we can get information regarding the GO term: name, *p-value* and number of gene sets it annotates (labeled "size"). While some BP annotations seem to be relevant (in orange/red), others in purple/dark blue from MF (*2'-5'-oligoadenylate synthetase activity* and *double-stranded RNA binding*) present low confidence levels (*i.e.*, high *p-values*).

As the M1.2 gene set is known to be involved at an early stage of the immune response [CB14], we further focused our study on that gene set. Clicking on it in the treemap view allowed to obtain more details on that particular gene set. From the indented tree view presented in figure 3.10, one can easily identify the five annotation terms computed by our pipeline. Among these annotation terms, three of them are from the BP ontology (with high levels of confidence) and two are from the MF ontology (with low levels of confidence). This provides a good cue on the biological role of the M1.2 gene set thanks to the three annotation terms: *signaling*, *response to virus* and *negative regulation of viral genome replication*.

With very few user interactions, we were able to retrieve the correct biological role of some gene sets and, in particular, those involved at the early stage of the immune response.

3.3 Conclusion

This chapter presented the importance of visualization facilities to help the interpretation of the functional annotation of gene sets. Three visualization prototypes were proposed in this chapter to explore results of the annotation of gene sets.

1. The two first visualizations were developed to graphically represent the results of an enrichment analysis pipeline used for annotating a single gene set.

2. The third visualization was developed to graphically represent the results of a pipeline combining enrichment and lexical similarity to annotate multiple gene sets.

The first analysis pipeline combines enrichment analysis, similarity measures, clustering and MICA terms. The results were displayed according to two visualization prototypes: a treemap showing the different clusters and the involved GO terms and a loom layout that represents the links existing between genes, GO terms and MICA terms. An important finding of this first study is that the visualization of MICA terms enables to summarize information. Thus, the overview visualization of the loom layout shows a clear synthesis of results through the MICA terms while the treemap requires that users explore one cluster after the other to get the involved GO terms. We also noticed that the enrichment analysis results in annotation terms having a high degree of redundancy and involves GO terms whose specificity may vary notably. For example, in section 3.1.3, `g:Profiler` provided very specific GO terms such as *response to type I interferon* as well as too generic GO terms, such as *response to stimulus* that corresponds to redundant information because it is an ancestor term of *response to type I interferon*. This issue is addressed in the next chapter.

The second pipeline combines enrichment, GO term alignment and simplification stages. The results were displayed within a visualization prototype called `MOTVIS`. It uses two interconnected views: a treemap view that provides an overview but also displays detailed information about gene sets, and an indented tree view that enables to focus on annotation terms of interest. Important findings from this work are the following: (i) the lexical analysis is useful to remove redundancy between terms describing the same notion across multiple knowledge resources, and (ii) the use of the GO structure within the visualization (even if the DAG was simplified into a tree) enables to reduce the visual complexity and thus facilitates the interpretation of results. Additionally, including the gene sets as leaves in the tree allows to remove potential redundancies between related GO terms associated to the same gene set. For that, we kept only the association with the most informative GO term. We finally illustrated the efficiency of this pipeline with a case study on immune response data.

The lexical analysis performed by `OntoEnrich` was useful to map (completely and partially) terms coming from other knowledge resources than GO. However, some knowledge resources describe very different information from GO, making it impossible to map the terms. For example, the resources used by `g:Profiler` like HPO or TRANSFAC that do not contain any term which can be mapped lexically to GO terms (except for some occasional partial matches). For this reason, the lexical analysis may not be well adapted to establish mappings between terms from knowledge resources describing distinct but complementary notions. We tried to deal with this issue and present our preliminary results in chapter 5.

Chapter 4

Annotating gene sets by using semantic similarity measures

Over the past decade, the revolution in new sequencing technologies has strongly supported the production of omics data with the aim to improve our understanding of the relations between genotype and phenotype. These produced data involved the need to analyze the gene sets in order to identify their biological function, and then to synthesize their key annotation information to help biologists with their interpretation.

In this frame, many tools have been developed to support gene set analysis and the visualization of their annotations. Most of these tools are based on statistical enrichment methods that usually involve two stages:

- an *a priori* stage that aims to synthesize the annotation by selecting the over-represented terms.
- an *a posteriori* stage which removes the potentially redundant information by using the structure of the knowledge resource from which the annotation terms come.

Moreover, as commented in section 2.4, the statistical-based methods tend to highlight the most studied genes at the detriment of the poorly annotated genes during the analysis [BLG15; HTK18; Tom+18]. As a consequence, some genes do not appear in results, thus resulting in a loss of information.

In this chapter, we address such issues by proposing an original method that annotates gene sets by using semantic similarity measures. The structure of this chapter is as follows: in section 4.1, we introduce some work realized to provide an improvement of the gene set annotation by using semantic similarity and present the motivation of this chapter. In section 4.2, we evaluate the impact of nine semantic similarity measures on a new method proposed to annotate gene sets with *Gene Ontology* (GO) terms. In section 4.3, we describe an improved version of this method by adding a last step that provides a synthetic annotation and we present the web server `GSAn` as well as its implementation as an R package (`RGSAn`). We conclude this chapter in section 4.4.

4.1 Gene set annotation with semantic similarity measures

4.1.1 State of the art of alternatives to enrichment analysis

As described in section 2.4 and in chapter 3, the challenges that are faced by enrichment analysis are: (i) a large number of redundancies among annotation terms, and (ii) many missing genes deemed to be poorly annotated. The semantic similarity has been considered as a solution to reduce the over-represented GO terms by making use of external and/or internal information from the ontology for comparing GO terms and identifying similarities between them. Xu et al. [Xu+09] proposed an *a priori* step that clusters similar GO terms based on their semantic similarity and then used them for an enrichment analysis. Supek et al. [Sup+11] developed the `REVIGO` web service with the aim to remove redundancies obtained by enrichment analysis tools. `ClusterProfiler` recently included an *a posteriori* step for reducing redundancies by using semantic similarity measures [Yu18]. However, even if these tools succeeded in reducing the redundancy after the enrichment analysis, they did not solve the problem of missing genes.

Another solution is to directly use alternative approaches of enrichment analysis that can take advantage of the ontology organization where terms are hierarchically structured according to the granularity of information. These approaches, designated as *gene functional similarity*, are based on semantic similarity measures and aim to compare two genes according to their annotation terms [Zha+06]. The literature has given a variety of tools dedicated to the gene-to-gene analysis in order to group the genes sharing similar annotations, including `GOSim` [Frö+07], `GOSemSim` [Yu+10], `G-SESAME` [Du+09], `GFSAT` [Xu+13] and `GOGO` [ZW18], which differ from each other, among other things, by the strategy they adopt for calculating the similarity. These approaches have the advantage to compute a comparison score between terms and the extension of such computation to deal with a gene set (that may contain a large number of genes) is certainly of great interest to synthesize the gene set functional information. Nevertheless, these methods can group similar annotation terms but they do not propose a strategy to synthesize such information.

4.1.2 Motivation to develop a method based on semantic similarity measures for annotating gene sets

This chapter aims to describe an alternative approach to enrichment analysis in order to compute a synthetic annotation for a given gene set by using semantic similarity measures that reduce *a priori* the large number of annotation terms. This strategy would facilitate the biological interpretation of a gene set that may contain hundreds of genes. To the best of our knowledge, despite the high number of semantic similarity measures that exist to compare two terms, no work has been proposed to evaluate the impact of using a given semantic similarity measure rather than another measure. In this frame, we consider that a relevant gene set annotation needs to meet specific features for providing relevant information to domain experts. We propose to evaluate the impact of using different measures while considering the following features that define the relevant criteria for a "good" synthetic gene set annotation [BPG16]:

- The number of annotation terms has to be drastically reduced, while the relevant terms that representatively annotate the gene set must be retained (designated as **synthesis**).

- The number of genes described by the selected terms (designated as **coverage**) has to be maximized.

Finding the best possible compromise between these two features is not an easy task while attempting to maintain a **sufficient level of details** supplied by the selected terms. In this context, we proposed a new method to evaluate the impact of semantic similarity measures in annotating gene sets.

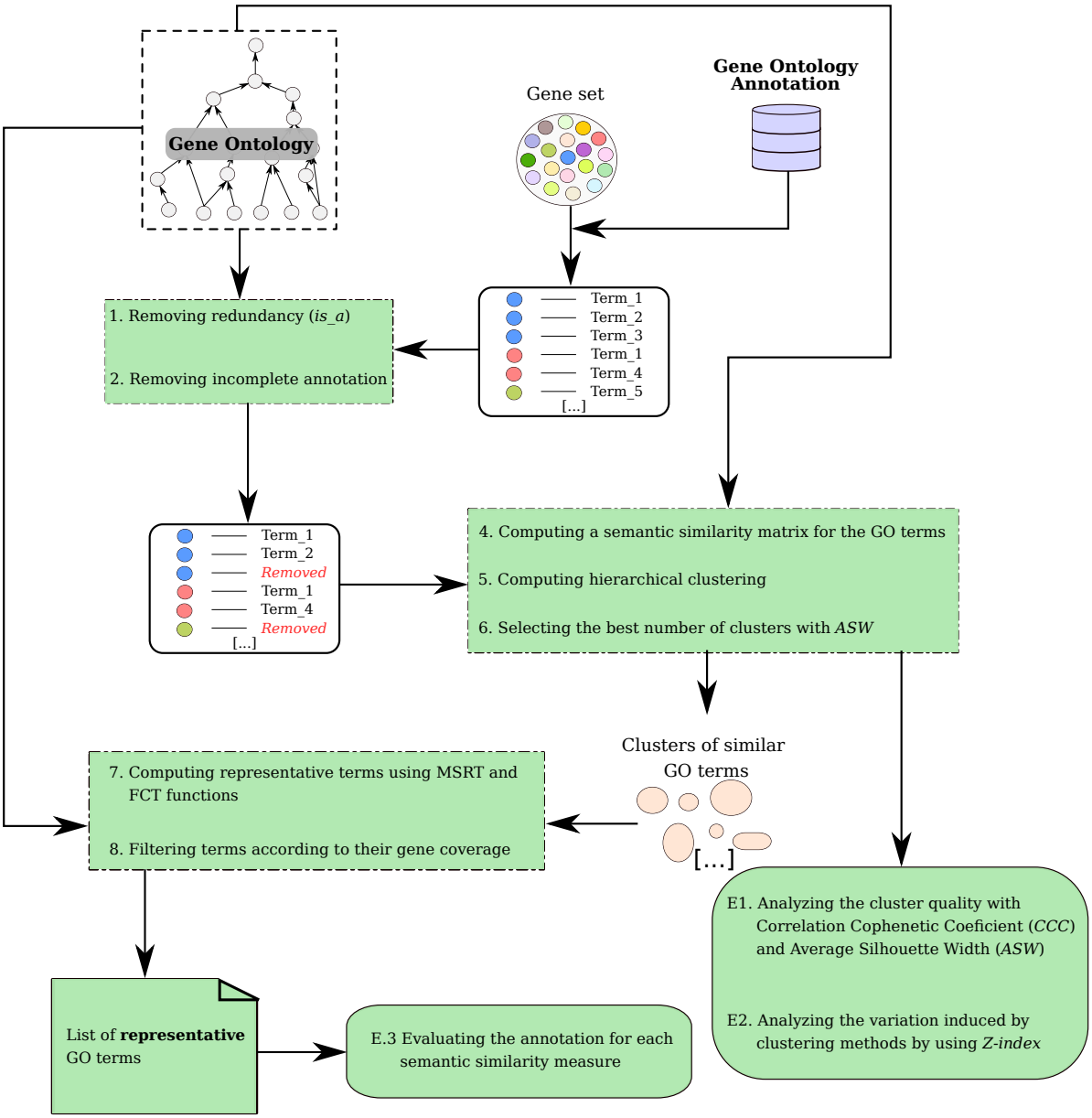


FIGURE 4.1: Proposed workflow to annotate gene sets and to study the impact of using different semantic similarity measures. The green dotted rectangles correspond to the steps implemented to annotate gene sets while the rounded rectangles correspond to the evaluation steps.

4.2 Impact of semantic similarity in the gene set annotation

We implemented a workflow in order to study the impact of using different semantic similarity measures while interpreting the gene sets (figure 4.1). First, we eliminated redundancy

and incompleteness among the GO annotation terms before computing semantic similarity matrices of pairs of terms. Secondly, we assessed the ability of each semantic similarity measure to compute the best partitions of the annotation terms by evaluating the clustering partition fitness and estimating the impact of varying hierarchical clustering methods. Finally, we examined the effectiveness of each semantic similarity measure in: (i) reducing the number of annotation terms while selecting the most representative terms of the investigated gene set, and (ii) providing annotation for the maximum number of genes included in this gene set.

4.2.1 Semantic similarity measures

Features of GO terms

As mentioned in section 2.3.2, each GO term has many features that can be used to quantify its similarity with other GO terms. Focusing on the *Information Content* (IC) feature, two families of methods were presented: *annotation* (or *extrinsic* methods) and *topology-based* (or *intrinsic* methods). Most of existing approaches are computing the *IC* in a similar way [MCM17], as follows:

$$IC(T) = -log(p(T)) \tag{4.1}$$

where $p(T)$ corresponds to the probability of encountering a term T within the taxonomy structure or external annotations. The probability p is monotonic, which means that for two related terms in a taxonomy, *i.e.*, T_1 *is_a* T_2, the probability of encountering T_1 is equal or inferior to the probability of encountering T_2. Thus, if the studied taxonomy has a unique top node, then its probability is 1 and its corresponding *IC* is 0 [Res95; JC97].

The *extrinsic* category corresponds to metrics that are using external knowledge. The first introduced and the most famous metric in this category is the one proposed by [Res95], which is based on the frequency of a concept (term) within a corpus. In the context of GO, this *IC* is related to the occurrence of a GO term within the *Gene Ontology Annotation* (GOA), as follows: the more frequently a term is used to annotate genes, the lower is its *IC* value. This assumption is based on the *true-path-rule* presented in section 2.2.1, stating that when a gene is associated with a given term, then it is also associated with any of its ancestors. Thus, the probability of encountering a term T based on the annotation resource is defined as follows:

$$p(T) = \frac{f(T) + \sum_{t \in children(T)} f(t)}{f(Root)} \tag{4.2}$$

where *f(T)* is the number of genes in a given organism annotated by the term T (the *Root* being *Biological Process* or BP, *Cellular Component* or CC or *Molecular Function* or MF). Thus, the *IC* proposed by Resnik [Res95] (or IC_R) is computed as the equation (4.1).

Subsequently, Seco et al. [SVH04] argued that the *IC* should be computed independently from external knowledge. More precisely, these authors claimed that the *IC* should be *intrinsic* to the ontology, meaning that only the knowledge contained within the ontology should be used to compute such a measure. In their metric, the probability of encountering a given term T

is defined as follows:

$$p(T) = \frac{|descendants(T)|}{|descendants(Root)|} \tag{4.3}$$

where *descendants(T)* is the set of descendants of the term T within GO. The IC of Seco et al. [SVH04] (or IC_S) is computed as follows:

$$IC_S(T) = 1 - \frac{log(|descendants(T)|+1)}{log(|descendants(Root)|)} \tag{4.4}$$

Thus, the IC is between 0 and 1, with the IC of 1 corresponding to the top term within the taxonomy, and the IC of 0 to the leaf terms.

However, as emphasized by Mazandu and Mulder [MM12a], this type of IC considers all descendants similarly regardless of their depth. Thus, it does not distinguish terms with different specificities. To address this issue, other *intrinsic ICs* have been proposed [MM13]. In particular, Mazandu and Mulder [MM12a] introduced an alternative *intrinsic IC* that considers the depth of the term T and its descendants, defined as follows:

$$p(T) = \begin{cases} 1 & \text{if } T \text{ is a root.} \\ \prod_{t \in ancestors(T)} \frac{p(t)}{|descendants(t)|} & \text{otherwise.} \end{cases} \tag{4.5}$$

where $ancestors(T)$ is the set of ancestors of the term T within GO. Thus, the IC of Mazandu (or IC_{GOu}) is also computed following the equation (4.1).

Apart from the IC, the following features of the two GO terms to be compared may also be considered to quantify their similarity: their distance (*i.e.*, the shortest path from one GO term to the other GO term) and their common ancestors. Concerning the latter, the following two main features have been introduced for distinguishing ancestors that might be more relevant than others: (i) the *Lowest Common Ancestor* (LCA) which is the common ancestor of the two GO terms that is the most specific, *i.e.*, having the maximal depth within the taxonomy, and (ii) the *Most Informative Common Ancestor* (MICA) which is the common ancestor having the highest IC value.

Investigated semantic similarity measures

Different classifications of semantic similarity measures were presented in section 2.3.2. The purpose of this section was not to propose a new categorization of existing semantic similarity measures for comparing GO terms but rather to evaluate their varying impact while analyzing gene sets. Thus, we selected nine pairwise semantic similarity measures according to the classifications provided by Pesquita et al. [Pes+09], Guzzi et al. [Guz+12] and Mazandu et al. [MCM17] by choosing at least one measure belonging to each category. The measures are listed below with a description of the feature(s) that they use.

- Ganesan [GGW03] adapted to the comparison of terms by Sanfilippo et al. [San+07]: the longest path from the root to both terms and their LCA;
- Leacock & Chodorow [LC98] normalized (LC): the shortest path between the two terms and the maximal depth within the ontology;

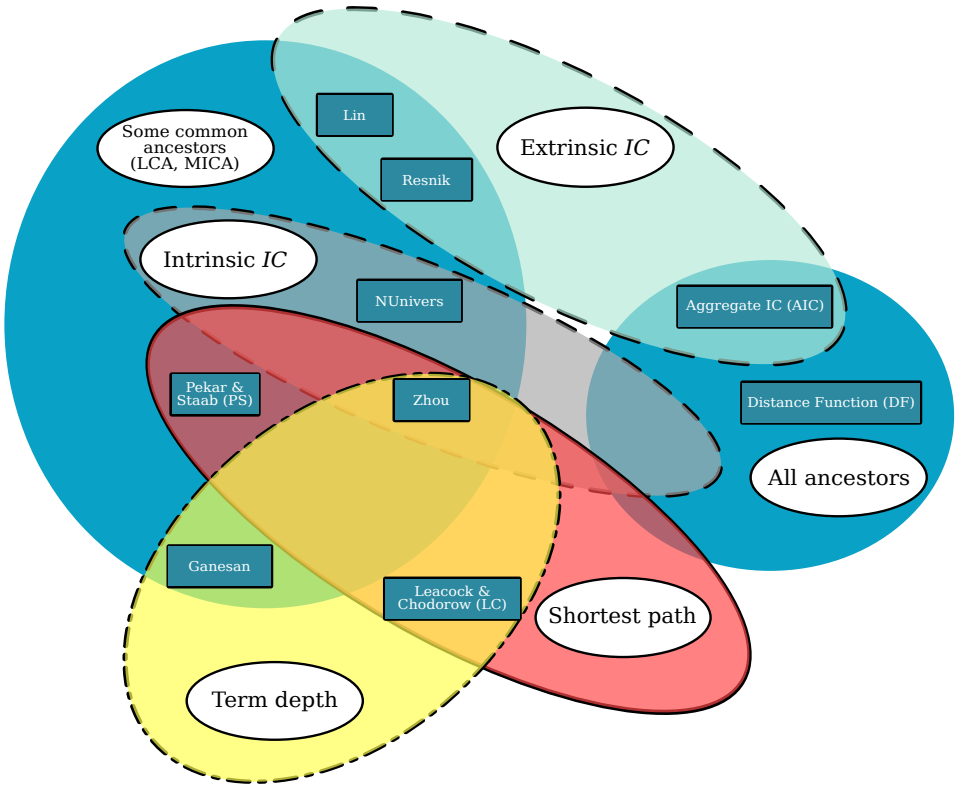


FIGURE 4.2: Classification of the nine semantic similarity measures (represented in *dark blue* rectangles) investigated in this study according to the features that they use (adaptation from Guzzi et al. [Guz+12]).

- Pekar & Staab (PS) [PS02]: the shortest path between the two terms and the shortest path between their LCA and the root term;
- Zhou [ZWG08b]: the IC_S of each term and of their MICA as well as the shortest path between the two terms and the maximal depth within the ontology;
- Resnik [Res95] normalized according to Jain and Bader's approach [JB10]: the IC_R of the MICA of the two terms and the maximal value of the IC_R within GO;
- Lin [Lin98]: the IC_R of each term and of their MICA;
- Nunivers [MM12a]: the IC_{GOu} of each term as well as the IC_{GOu} of their MICA;
- Distance Function (DF) [Que+15b]: the ancestors of the two terms;
- Aggregate *IC* (AIC) [Son+14]: the IC_R of ancestors of the two terms.

As shown in table 4.1, we chose three *edge-based* measures, five *node-based* measures (including a *graph-based* measure) and one *hybrid* measure. We selected more *node-based* measures than *edge-based* measures because the latter were deemed less efficient in comparing GO terms [Pes+09]. Figure 4.2 was adapted from the categorization proposed by Guzzi et al. [Guz+12] to illustrate that the measures investigated in this study use all the usual features of GO terms. To consider the two previously introduced categories of *IC*, we replaced the "Term *IC*" feature used by Guzzi et al. [Guz+12] with the following two distinct features: "Extrinsic *IC*" (*i.e.*, the IC_R) and "Intrinsic *IC*" (*i.e.*, the IC_S and IC_{GOu}). Notably, we did not

choose any vector space model-based measure as these measures compare terms not only according to features specific to the terms within the ontology but also according to external knowledge (*i.e.*, the genes they annotate).

TABLE 4.1: The nine semantic similarity measures investigated in this study.

Name	Category	Measure
Ganesan [San+07]	Edge-based	$Sim_{Ganesan}(T_a,T_b) = \frac{2\cdot\delta(T_{lca})}{\delta(T_a)+\delta(T_b)}$
[LC98] normalized	Edge-based	$Sim_{LC}(T_a,T_b) = 1 - \frac{log(D_{sp}(T_a,T_b))}{log(2\cdot\delta_{max})}$
[PS02]	Edge-based	$\frac{D_{sp}(T_{lca},root)}{D_{sp}(T_{lca},root)+D_{sp}(T_a,T_b)}$
[ZWG08b]	Hybrid	$Sim_{Zhou}(T_a,T_b) = 1-k\cdot\frac{log(D_{sp}(T_a,T_b)+1)}{log(2\cdot\delta_{max}-1)} - (1-k)\cdot\frac{IC_S(T_a)+IC_S(T_b)-2\cdot IC_S(T_{mica})}{2}$
Resnik normalized [JB10]	Node-based	$Sim_{Resnik}(T_a,T_b) = \frac{IC_R(T_{mica})}{IC_{R_{max}}}$
[Lin98]	Node-based	$Sim_{Lin}(T_a,T_b) = \frac{2\cdot IC_R(T_{mica})}{IC_R(T_a)+IC_R(T_b)}$
Nunivers [MM12a]	Node-based	$Sim_{Nunivers}(T_a,T_b) = \frac{IC_{GOu}(T_{mica})}{max\{IC_{GOu}(T_a),IC_{GOu}(T_b)\}}$
Distance Function [Que+15b]	Edge-based	$Sim_{DF}(T_a,T_b) = \frac{\lvert ancestors(T_a)\cap ancestors(T_b)\rvert}{\lvert ancestors(T_a)\cup ancestors(T_b)\rvert}$
Aggregate *IC* [Son+14]	Node-based/Graph-based	$Sim_{AIC}(T_a,T_b) = \frac{\sum_{t\in ancestors(T_a)\cap ancestors(T_b)} 2\cdot SW(t)}{\sum_{t_a\in ancestors(T_a)} SW(t_a) + \sum_{t_b\in ancestors(T_b)} SW(t_b)}; SW(x) = \frac{1}{1+e^{-\frac{1}{IC_R(x)}}}$

$\delta_{\mathbf{max}}$: maximal depth of the ontology; $\delta(\mathbf{T_x})$: longest path between T_x and the root of the ontology; $\mathbf{T_{lca}}$: lowest common ancestor term; $\mathbf{T_{mica}}$: most informative common ancestor term; $\mathbf{D_{sp}}(\mathbf{T_x},\mathbf{T_y})$: the shortest; path distance between T_x and T_y; $\mathbf{IC_S}$: information content of Seco et al. [SVH04]; $\mathbf{IC_R}$: information content of Resnik [Res95]; $\mathbf{IC_{GOu}}$: information content of Mazandu and Mulder [MM12a]; **SW**: semantic weight by Song et al. [Son+14]; **k**: contributor factor, which can be adapted manually.

4.2.2 Selection of Gene Ontology terms

Recovering the annotation

We used GO due to the existence of a hierarchical structure of terms and a rich annotation of genes for multiple organisms from GOA (these resources have been described in detail in section 2.2.1). To deal with GO, we also use the OWL format. For GOA, the used format was the *Gene Association File* (GAF) 2.1.

To illustrate and discuss the results of this work, we applied our methods on [B-346] and [C-260] datasets (presented in section 1.4), whose gene sets are related to immune response.

Eliminating the inappropriate annotations

First, we removed the **inappropriate annotations**. An inappropriate annotation is defined by any association where the GO term does not provide relevant information. An annotation can be inappropriate for two reasons: **redundancy** and **incompleteness**.

The first stage of reduction was easily applied by eliminating **redundancy** [Jan+11] among the GO terms that annotate each gene in a given set. This stage consists in eliminating annotations that involve a GO term if at least one of its descendant terms annotate the same gene. To do so, when GO terms annotating a given gene are hierarchically-related (*i.e.*, one GO term is a parent or, more generally, an ancestor of another GO term), we only kept the most specific GO term because it provides more precise information.

The second stage aimed to eliminate incomplete annotations. According to the definition given by Faria et al. [Far+12], **incompleteness** corresponds to cases where the function of a gene is not fully described. Among the existing GO annotations, authors have emphasized that GO terms that are too generic constitute an incomplete annotation. Thus, we decided to remove such GO terms. To identify these incomplete terms, we used an alternative measure to the one proposed by Faria et al. [Far+12], who considered GO terms with more than 10 descendants as incomplete annotations. In practice, we computed the IC_{GOu} distribution of the GO terms used in GOA human. For that, for a particular GO term, its IC_{GOu} is included in the distribution as many times as there are genes associated with this term. For example, the GO term *T cell receptor signaling pathway* (`GO:0050852`) is associated with 478 genes and has an IC_{GOu} of 220.73. This way, the IC_{GOu} is considered 478 times for the distribution. Then, we removed the GO terms whose IC_{GOu} was in the first quartile of such distribution, which we considered to be incomplete annotations.

4.2.3 Clustering annotation terms using computed semantic similarity matrices

According to Thomas [Tho17], biological processes represent the objectives that an organism is "programmed" to realize. In addition, most specific annotation terms (*i.e.*, with a depth over 5) belong to the BP ontology (figure 4.3). Based on these considerations, we reduced the scope of our analysis by strictly focusing on the GO terms in this ontology since they provide a clear interpretation of the roles of genes.

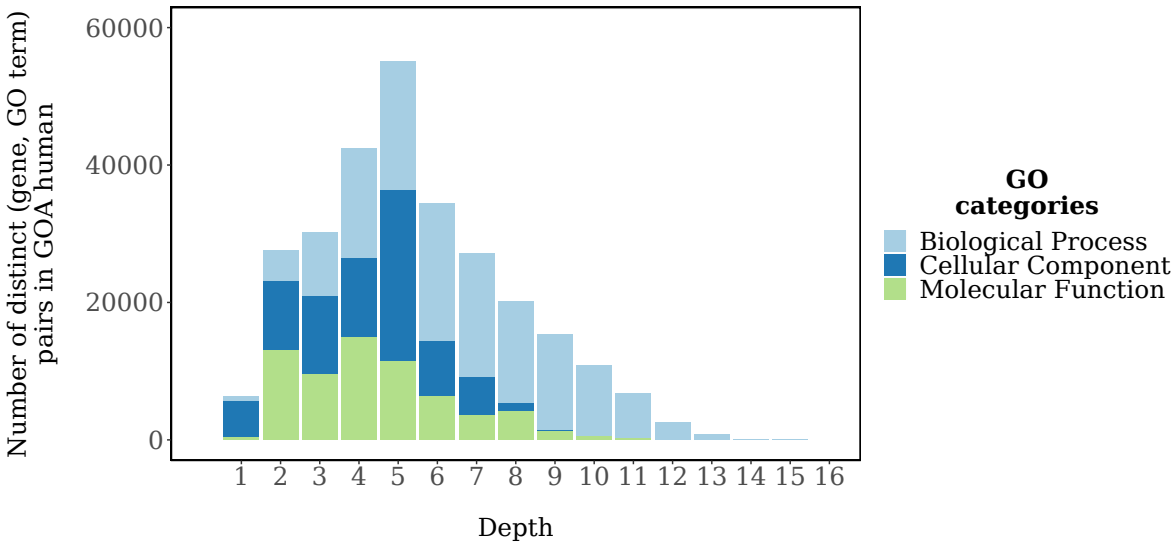


FIGURE 4.3: Depth of GO terms in the annotations provided by GOA human.

For each of the nine measures displayed in table 4.1, we computed matrices containing the semantic similarity value of each pair of GO terms, corresponding to biological processes that are part of the investigated gene set. Then, we applied a clustering method in order to group GO terms from the semantic similarity matrix into subgroups in such a way that similar GO terms are grouped together [MR05]. Multiple types of clustering methods might be able to provide groups of similar terms. These clustering methods were separated into two main groups [JMF99]:

1. *Hierarchical clustering*: Clusters are formed by iteratively dividing the elements of the matrix using a *top-down* (divisive) or a *bottom up* (agglomerative) approach. This type of approach generates a hierarchy tree of similar elements into a structure called dendrogram.

2. *Partitional clustering*: Clusters are assigned without producing any hierarchical clustering. Instead, this type of clustering optimizes a given criterion function [LW14].

In this work, we decided to use agglomerative hierarchical clustering methods to compute clusters of terms [BGL11]. The choice of using such unsupervised hierarchical clustering methods was motivated by the existence of hierarchical relations that connect the GO annotation terms represented within a DAG [SSZ04]. Using the list of pertinent features provided by Hennig [Hen15], the following three characteristics were required to define "good" clusters: (i) the within-cluster dissimilarity had to be small, (ii) the between-cluster dissimilarity had to be large, and (iii) the terms in a cluster had to be well-represented by the centroid or a very small group of terms (to help the identification of synthetic and relevant terms).

The most classical hierarchical clustering methods differ in how they define the perimeter of the resulting clusters when adding a new term. In the *Single Linkage Hierarchical Method* (SLHM), the distance between two clusters is given by the shortest distance that can be calculated between two terms from these two clusters. In the *Complete Linkage Hierarchical Method* (CLHM), this distance corresponds to the longest distance between two terms. Finally, the *Average Linkage Hierarchical Method* (ALHM) uses the average distance between each term from each cluster.

Even if no relation exist among the data, hierarchical clustering techniques always create partitions [HKK05]. Consequently, assessing the relevance of the resulting partitions is essential. There are different validation measures that have been categorized according to three criteria [HBV01]:

- *External* criteria evaluate the clustering results with an *a priori* knowledge about what is expected.

- *Internal* criteria evaluate the inner structures of the data of the clustering results themselves in order to assess the quality of the clustering (*i.e.*, *Cophenetic Correlation Coefficient* or *CCC* metric).

- *Relative* criteria evaluate the relation between *compactness* and *separation* of a given clustering result. In that case, these criteria evaluate the partitions obtained from the clustering rather than the clustering algorithm itself (*i.e.*, *Average Silhouette Width* ou *ASW*).

Due to the fact that we do not have an *a priori* knowledge to compare the clustering results, only *internal* and *relative* criteria were used.

Additionally, based on the assumption that the best semantic similarity measure should be able to reproduce the same term partition while varying the hierarchical clustering methods, we were interested in providing additional criteria to compare partitions. In the literature, there are many measures providing a comparison between two partitions for the same

dataset [FM83; RWR05; Mei07; YS10]. Nevertheless, measures comparing two partitions including the dendrogram structure are less frequent [MZ12]. Using dendrogram structures allows to observe differences in the organization of data after applying two clustering methods. Therefore, a good semantic similarity measure should be able to represent close dendrograms while varying the hierarchical clustering methods.

Analyzing the cluster quality with *CCC* and *ASW*

To test the ability of the clustering approaches to fit well the data, we computed the *internal criterion CCC* [SR62] of the resulting dendrograms. This internal metric aims to measure how the original pairwise distance between terms (given by semantic similarity measures) is retained within the computed dendrogram. The *CCC* score between the original term pairwise distance matrix Y and the dendrogram Z is calculated as follows:

$$CCC(Y,Z) = \frac{\sum_{i<j}(Y_{ij}-y)(Z_{ij}-z)}{\sqrt{\sum_{i<j}(Y_{ij}-y)^2 \sum_{i<j}(Z_{ij}-z)^2}} \tag{4.6}$$

where Y_{ij} is the distance between terms i and j given by the original pairwise matrix Y, and Z_{ij} is the distance between i and j computed within the dendrogram Z. The y and z values refer to the average distances within Y and Z, respectively.

Thus, we used the three clustering methods mentioned above while varying the semantic similarity measures. This score aims to compare the distances between the data: (i) within the dendrogram, and (ii) within the original semantic similarity matrix. The resulting coefficient corresponds to the square of the coefficient of determination and indicates the proportion of variance explained by the clustering results. Thus, a value close to 1 reflects a perfect correspondence.

As shown in figure 4.4, we observe that the *CCC* scores given by the SLHM, CLHM and ALHM methods follow the same trend, as they tend to increase or decrease according to the semantic similarity measures. However, there is a significant difference in the *CCC* score dispersion among the clustering methods. We observe that the *CCC* scores are noticeably lower with the SLHM method (*e.g.*, the *CCC* median scores of the LC and Zhou measures are below 0.50). The observation of such low varying degrees of quality could be expected for SLHM because this clustering method is known to often perform consecutive additions of terms and is more often used to reveal gradients in a dataset. The dispersion of scores according to the gene sets and semantic similarity was less important for the other two clustering methods, and in both cases, the *CCC* median scores are the highest considering similar measures, *i.e.*, DF and Nunivers. Notably, the Nunivers measure has the greatest *CCC* score with a very small dispersion according to the gene sets. In contrast, the measures that provide the most significant *CCC* score dispersion are not systematically the same according to CLHM and ALHM.

By using an unsupervised clustering method, which is a classical approach, we determined the optimal number of clusters based on the computation of a relative criterion: the *ASW* score [BGL11]. It consists in using the geometrical measures of cluster compactness and separation [VCH10], which are calculated while varying the number of clusters, *i.e.*, cutting the

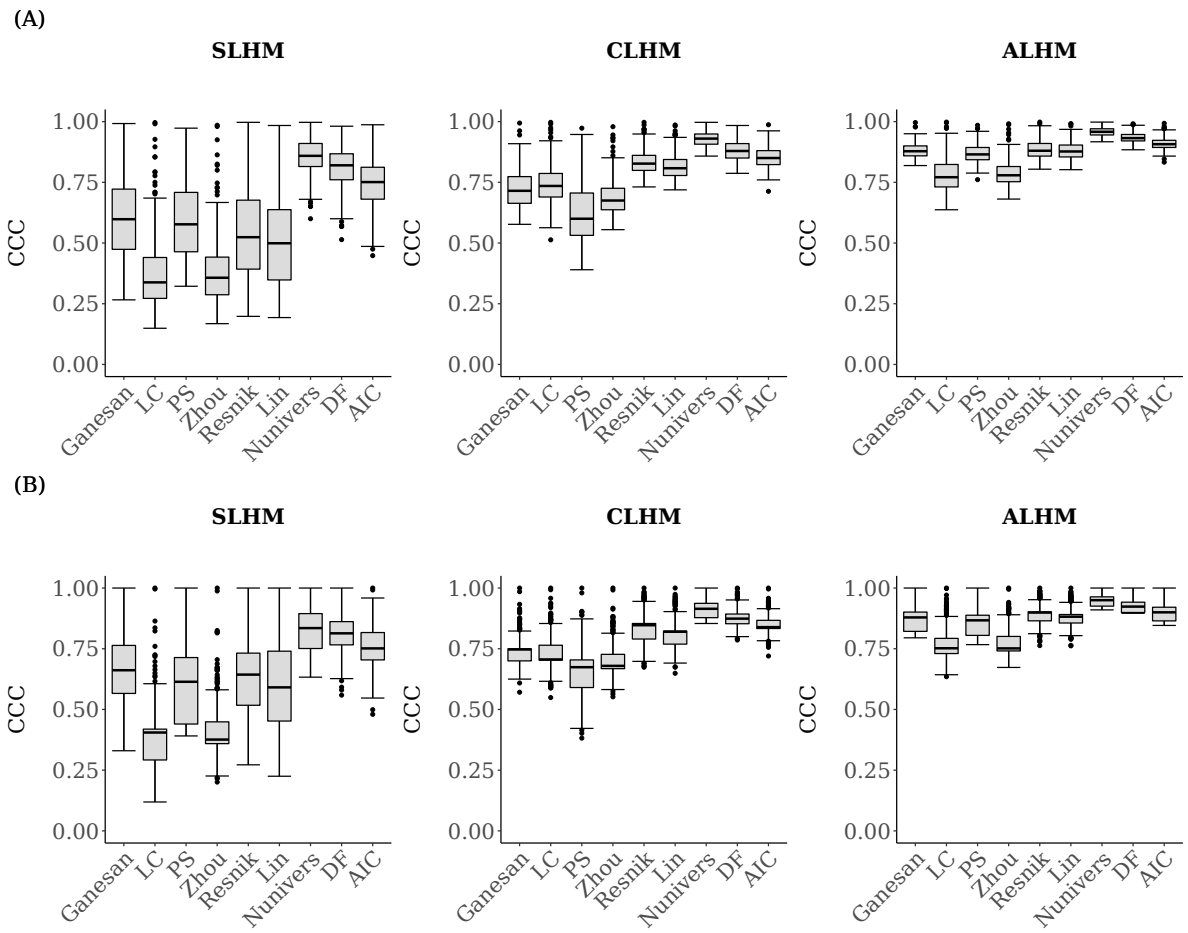


FIGURE 4.4: Cophenetic correlation coefficients (*CCC*) according to nine semantic similarity measures of the investigated gene sets: (A) for dataset [C-260] and (B) for dataset [B-346]. The following three *Linkage Hierarchical Methods* are presented: *Single* (SLHM), *Complete* (CLHM) and *Average* (ALHM).

dendrogram at different levels. For a given number K of clusters, we computed the average of the silhouette width. For each resulting cluster, the silhouette width is calculated as follows:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{4.7}$$

where i is a given term within a cluster, a_i is the average distance from i to all other terms within the cluster, and b_i is the minimum average distance from i to all other terms in any other cluster. Thus, the *ASW* score of the clusters for a given K is given by:

$$ASW_K = \frac{\sum_{i \in T} s_i}{|T|} \tag{4.8}$$

where T refers to the whole set of terms.

Then, considering the results obtained for each set of K clusters, the optimal partition (*i.e.*, the K value) can be deduced from the highest *ASW* score. The *ASW* score was then computed from the two datasets of gene sets to analyze its distribution. To guide the interpretation of the *ASW* score distributions, a score below 0.25 or above 0.50 might respectively correspond to an artificial or real structure within the data respectively [KR90]).

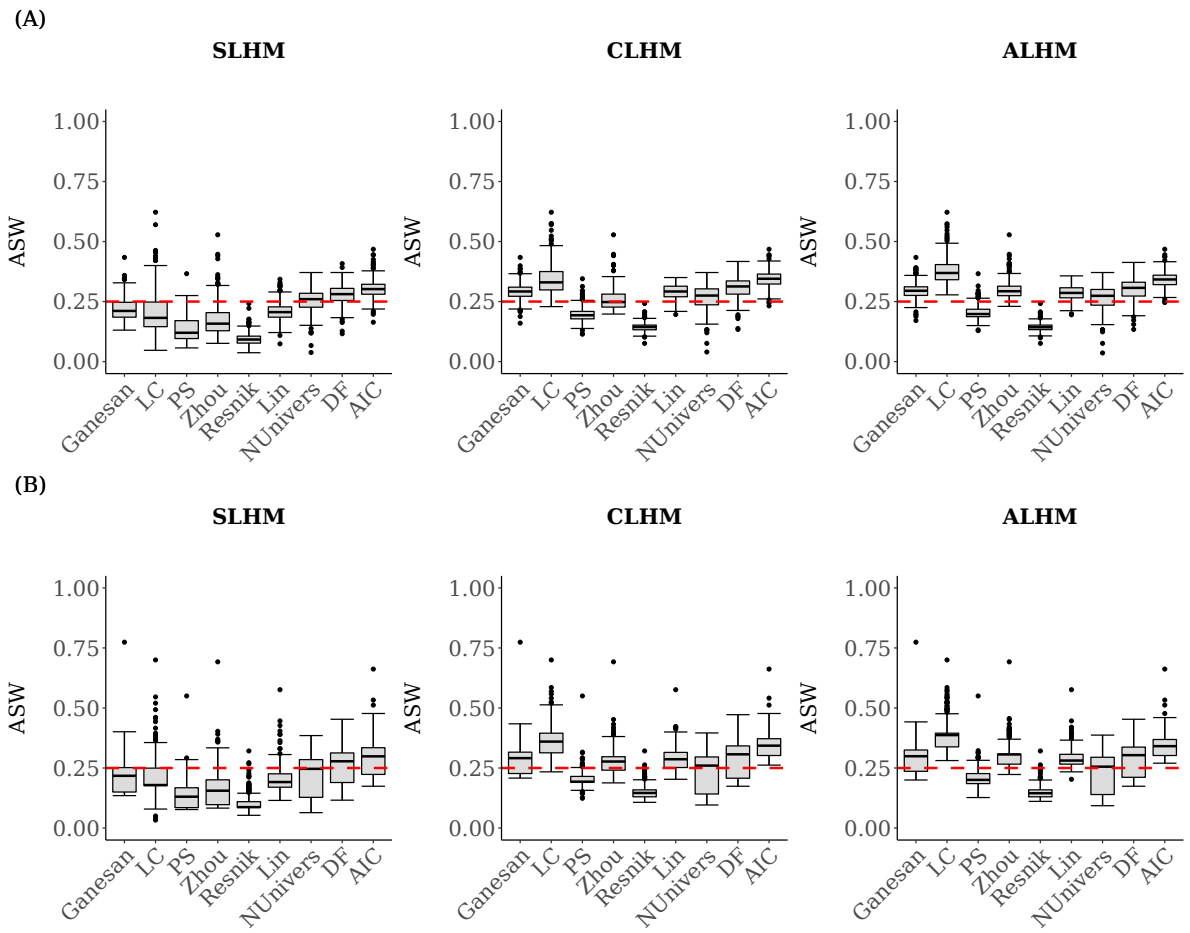


FIGURE 4.5: Average Silhouette Width (*ASW*) according to nine semantic similarity measures of the investigated gene sets: (A) for dataset [C-260] and (B) for dataset [B-346]. The following three *Linkage Hierarchical Methods* are presented: *single* (SLHM), *complete* (CLHM) and *average* (ALHM).

In order to evaluate the quality of the clusters of terms, we analyzed the *ASW* score for the three hierarchical clustering methods. This score was computed for each gene set in both datasets. As shown in figure 4.5, we observe that the *ASW* scores given by the CLHM and ALHM methods follow the same trend, while SLHM follows the same trend only for the semantic similarity measures NUnivers, DF and AIC (being part of the *node-based* category). That is an interesting point since these three semantic similarity measures have a close cluster quality for each hierarchical method, even if the *CCC* results for SLHM are the worst compared to the other hierarchical methods. Focusing on CLHM and ALHM (corresponding to better results with the *CCC* score), we can observe that the *ASW* score is below 0.25 for the PS and Resnik measures and between $0.26 - 0.50$ for the other measures. Thus, no semantic similarity measure managed to compute a global partition with a structuring score above 0.50, which is the threshold used to separate unstructured from structured clusters. This observation must be moderated by the specificity of the data in which an important number of genes are still unknown in some gene sets (as represented in the two figures in appendix A in which the coverage of the annotated genes can vary significantly depending on the gene sets) and, consequently, badly annotated [HTK18; Tom+18].

While the best clustering quality score with *CCC* was obtained by using the ALHM method and similar *ASW* scores with ALHM and CLHM methods, the following question emerges: for a given semantic similarity measure and a given gene set, *do we retrieve the same partitions of terms regardless of the clustering method used?* Thus, *how similar are the partitions given by two clustering methods for a given gene set and a given semantic similarity measure?* If the partitions are highly similars, this information could be considered as a reliability and robustness criterion for the semantic similarity measure due to the small influence of the clustering methods.

Variations induced by clustering methods

Based on the results obtained by the SLHM method in terms of *CCC* and *ASW* score, we did not use it to evaluate the variation between clustering methods because of its ineffectiveness in clustering data showing hierarchical relations (as shown in figure 4.4 and figure 4.5). Thus, we compared the partitions given by CLHM and ALHM using the *Z-index* [MZ12]. This value gives a criterion for evaluating the differences between two clustering methods for a fixed semantic similarity measure. This *Z-index* is computed as follows:

$$Z = \sum_k Z_k = \sum_k \sum_i \frac{|x_{1ik} - x_{2ik}|}{|x_{1ik}| + |x_{2ik}|} \tag{4.9}$$

where x_{1ik} and x_{2ik} indicate whether a pair ik of terms are within the same cluster in dendrograms X_1 and X_2. All potential partitions are simultaneously considered by varying k. The Z value provides the distance between two clustering methods by computing the difference in the status (*i.e.*, whether both terms are in the same cluster) of each pair at each stage of the procedure (by varying the k value).

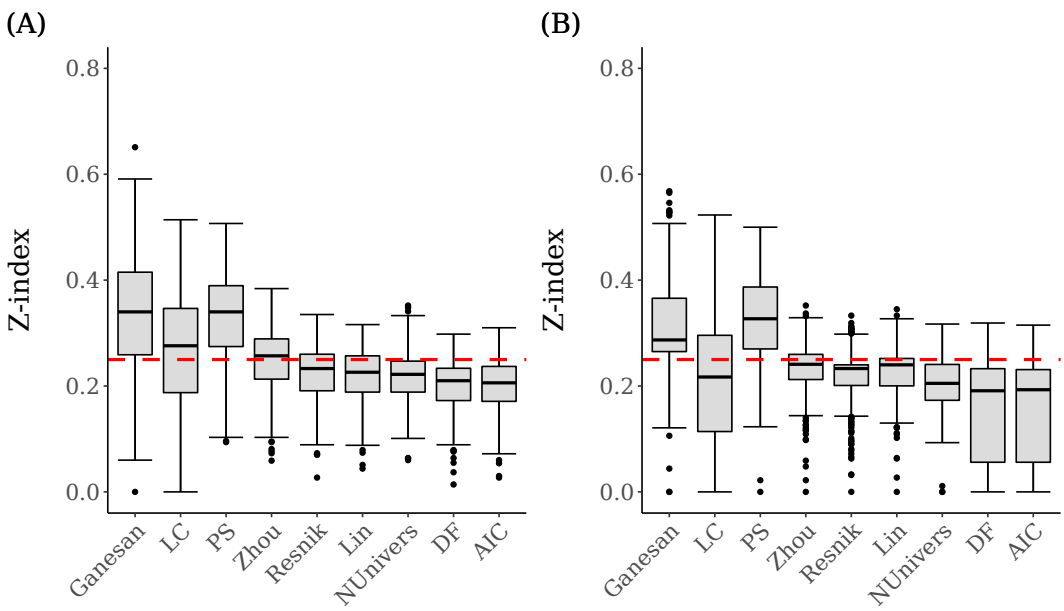


FIGURE 4.6: *Z-index* scores for comparing the whole structure of the dendrograms using CLHM and ALHM for the investigated gene sets: (A) for dataset [C-260] and (B) for dataset [B-346].

Thus, the smaller the differences between the results provided by both clustering methods, the more robust the semantic similarity measure is regardless of the clustering methods. This

measure assesses the possible combinations of all pairs of terms, which could be interpreted as a comparison of the intermediate structures given by both dendrograms before obtaining the different partitions (without considering *a priori* a given K number of optimal partitions). Then, this criterion can be considered to be a generalization of the classical metrics used to compare two partitions [MZ12] and it provides a score ranging from 0 to 1 that corresponds to a perfect resemblance and dissemblance, respectively, between the two dendrograms generated by each clustering method.

We used the *Z-index* to compare and decipher the similarities and differences between both partitions given by the CLHM and ALHM methods (figure 4.6). Based on the assumption that a pertinent semantic similarity measure should provide the same clusters regardless of the adopted clustering approach, we assume that the smaller the *Z-index* score is, the greater the similarity between two distinct clustering method partitions is. Thus, most semantic similarity measures gave similar results with *Z-index* scores smaller than 0.25, except for the Ganesan and PS measures on both datasets (LC and Zhou were also below 0.25 when computed on the [C-260] dataset).

As an intermediate conclusion, based on this analysis, we highlighted five semantic similarity measures (LC, Zhou, Resnik, Lin, Nunivers, DF and AIC) for which the choice of CLHM or ALHM did not show a significant impact regardless of the dataset. However, clustering using ALHM provided better *CCC* (figure 4.4) and *ASW* (figure 4.5) scores and was thus used for further analysis.

4.2.4 Identifying the most relevant representative terms

To provide a meaningful analysis, we assessed the efficiency of the semantic similarity measures in inferring synthetic and relevant terms by using the resulting partitions. Identifying the best trade-off between summarizing the information and preserving a good level of precision (given by the depth of terms) was challenging. Thus, first, only some terms were selected for each cluster, and secondly, the resulting clusters that annotated only a few genes were not included. Then, we compared the semantic similarity measures based on their capacity to provide: (i) a synthetic annotation with relevant terms, and (ii) a good coverage of the gene set while finally guaranteeing a fine-grained annotation.

The clustering method generates clusters of annotation terms that exhibit a certain level of similarity. The purpose of this step was to identify the most synthetic and pertinent GO terms, which are subsequently designated as **representative terms**, of each cluster to summarize their annotation information. The algorithm proposed for identifying such representative terms is provided in the next sub-section.

Description of the algorithm for identifying representative terms.

To identify the representative terms of a cluster, we developed a new algorithm based on two functions: *MSRT* (*Most Specific Representative Terms*) and *FCT* (*Find Combined Terms*). All terms are represented by a bitset that is set at the size of the cluster and initialized with all zeros. Bitsets are used to compare terms while traversing the ontology (figure 4.7A). Each position

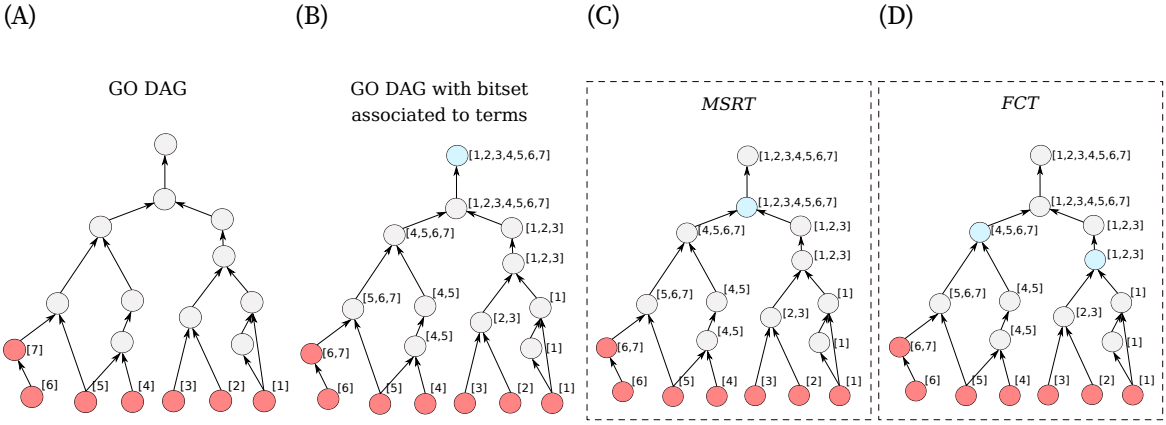


FIGURE 4.7: Steps for identifying the representative terms given a cluster of GO terms. (A) The GO DAG where the nodes stand for bitset to represent the use of terms within a cluster (colored in red). (B) The bitsets of ancestor terms of the GO terms within the cluster are filled following the *true-path-rule*. For example, the bitset $[5,6,7]$ stands for an ancestor term that has as descendants the terms 5, 6, and 7 of the given cluster. (C) From the top of the ontology, *MSRT* (*Most Specific Representative Terms*) searches a more specific GO term that involves all GO terms of the cluster. (D) From the representative GO term identified by *MSRT*, *FCT* (*Find Combined Terms*) identifies the potential combination of more specific terms whose bitset union satisfies the complete bitset.

of a bitset is associated with a term present in the original cluster, and, thanks to the *true-path-rule*, the position of each bitset associated with the ancestor terms of this term is set to 1 (figure 4.7B). Then, the bitset associated with a term summarizes if this term is hierarchically related to some terms of the cluster. In practice, this algorithm is as follows (see details in algorithm 1 and algorithm 2):

- The *MSRT* algorithm tries to identify for a given top term (initially set up to BP) a set of terms more specific in their biological meaning. Thus, the related bitsets are compared. If the bitsets are equal, the top term can be ignored and the corresponding descendant terms are retained as candidate representative terms (figure 4.7C).

- Then, the *FCT* algorithm is called for each candidate representative term returned by *MSRT*. As *MSRT*, the *FCT* algorithm compares bitsets to identify a combination of more specific terms related to the same terms of the cluster (figure 4.7D). This combination can vary from two to more combined terms by depending on the number of GO terms in the cluster. Thus, a number of combinations k for a given cluster is defined as follows:

$$f(c) = floor(\sqrt{|\frac{Nt}{10} - 1|}) + 2 \tag{4.10}$$

  where Nt is the number of terms in a given cluster c. To compute that, the minimal number of GO terms in the cluster must be over 5. Thus, for a cluster containing from 2 to 19 GO terms, the number of maximal combination is 2, from 20 to 49 GO terms, the number of maximal combination is 3, etc.

- At the end, we retain the combination(s) of representative terms having the best *IC* value. The *IC* of a combination is given by the mean *IC* of all terms in the combination.

Algorithm 1: MSRT(*top_term*, *GO*)

```
Input : top_term is a term,
        GO is an object with two attributes: (i) bitset(x) is the bitset associated to term x
        and (ii) children(x) is the list of terms that are children of x.
Ouput: representative_terms is the set of specific representative terms.
1  terms_to_visit := ∅; representative_terms := ∅; visited_terms := ∅
2  push(terms_to_visit, top_term)
3  while terms_to_visit ≠ ∅ do
4      candidate_term := pop(terms_to_visit)
5      if candidate_term ∉ visited_terms then
6          is_child_representative := false
7          push(visited_terms, candidate_term)
8          for child_term ∈ GO.children(candidate_term) do
9              if GO.bitset(child_term) == GO.bitset(top_term) then
10                 is_child_representative := true
11                 push(terms_to_visit, child_term)
12             end
13         end
14         if is_child_representative == false then
15             push(representative_terms, candidate_term)
16         end
17     end
18 end
19 return representative_terms
```

Algorithm 2: FCT(*representative_term*, *k*, *sct*, *GO*)

```
Input : representative_term is a term representing the cluster,
        k is the maximum number of combinations,
        sct is the minimum number of bits shared by terms,
        GO is an object with two attributes: (i) bitset(x) is the bitset associated to term x
        and (ii) children(x) is the list of terms that are children of x.
Ouput: representative_terms_set is a set of sets of most specific representative terms.
1  if k > size(GO.children(representative_term)) then
2      k := size(GO.children(representative_term))
3  end
4  combinations_set ← all combinations of k terms from GO.children(representative_term)
5  representative_combination_found := false
6  for combination ∈ combinations_set do
7      GO.bitset(combination) ← union of the bitsets of terms part of combination
8      number_of_shared_bits ← number of bits shared by terms part of combination
9      if GO.bitset(combination) == GO.bitset(representative_term) and
10        number_of_shared_bits < sct then
11         combined_representative_terms = ∅
12         for term_in_combination ∈ combination do
13             push(combined_representative_terms, MSRT(term_in_combination, GO))
14         end
15         push(representative_terms_set, combined_representative_terms)
16         representative_combination_found := true
17     end
18 end
19 if representative_combination_found == false then
20     push(representative_terms_set, {representative_term})
21 end
22 return representative_terms_set
```

Filtering representative terms according to gene coverage

As the representative terms are conditioned by the size of the clusters, one can obtain a very interesting representative term while the number of genes described by this term may not

be sufficient. For that, only representative terms that annotate a minimum of three genes in the investigated gene sets were retained, designated as **synthetic terms** throughout the remainder of this chapter. We applied this filter because our aim was to obtain a synthetic annotation at the end of the proposed method.

Considering the criteria used to define a "good" synthetic gene set annotation, we addressed the following key questions: *what is the level of details given by the remaining representative terms?* and *what is the number of genes that can be related with these representative terms?*

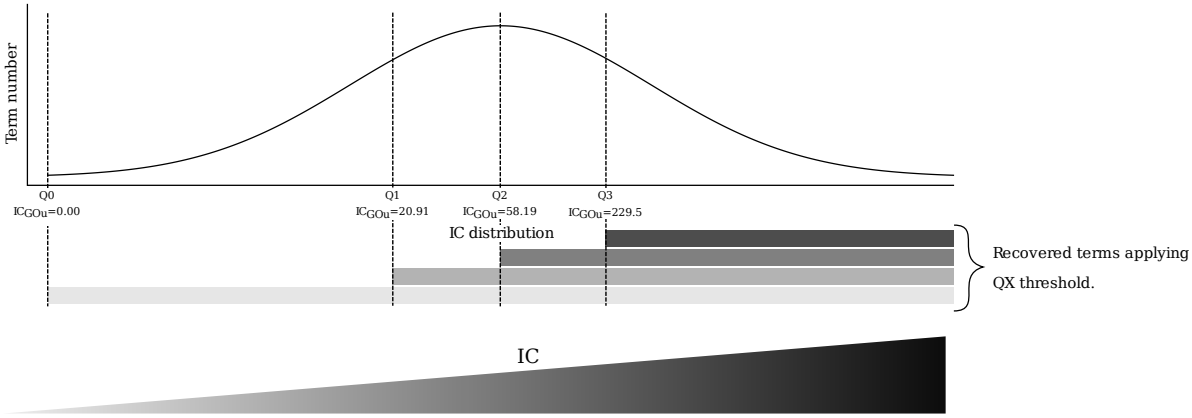


FIGURE 4.8: IC_{GOu} distribution of GO terms used in GOA human.

Evaluating the annotation for each semantic similarity measure

We compared the impacts of the semantic similarity measures by examining the representative terms obtained by each measure. As previously stated in the introduction, a suitable synthetic gene set annotation reduces the number of terms while maintaining a sufficient level of details within the synthetic terms. Based on the distribution given by the computed IC_{GOu} values of the whole set of terms used in GOA human, we divided the value range into four contiguous intervals with an equal density of terms (figure 4.8). This categorization of the results was used to assess the semantic similarity measures' impacts on: (i) the number of synthetic terms selected to annotate the genes, and (ii) their level of details. Furthermore, the distinction between two measures that equally decrease the number of terms or are scored with the same number of related genes had to be analyzed in depth according to the level of details given by each group of terms. In practice, the cumulative percentage of synthetic terms was computed for each quartile of the IC_{GOu} values by simultaneously considering the term and gene coverage sides.

Then, to assess the impact induced by each semantic similarity measure, we first evaluated the reduction in the number of terms between the original set of annotation terms and the filtered representative terms. As shown in figure 4.9, we observed a subsequent drastic decrease in the number of terms using all semantic similarity measures. In particular, the most striking decrease is observed with the LC and Zhou measures, where the remaining number of terms is very small.

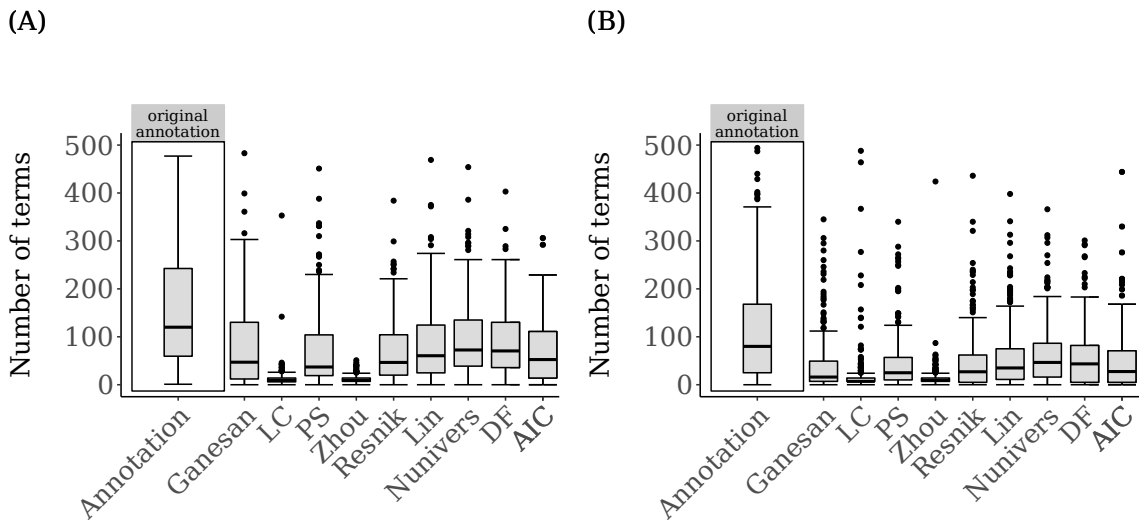


FIGURE 4.9: Number of representative terms obtained after applying each semantic similarity measure for the investigated gene sets: (A) for dataset [C-260] and (B) for dataset [B-346]. The "Original annotation" box-plot corresponds to the initial number of annotation terms.

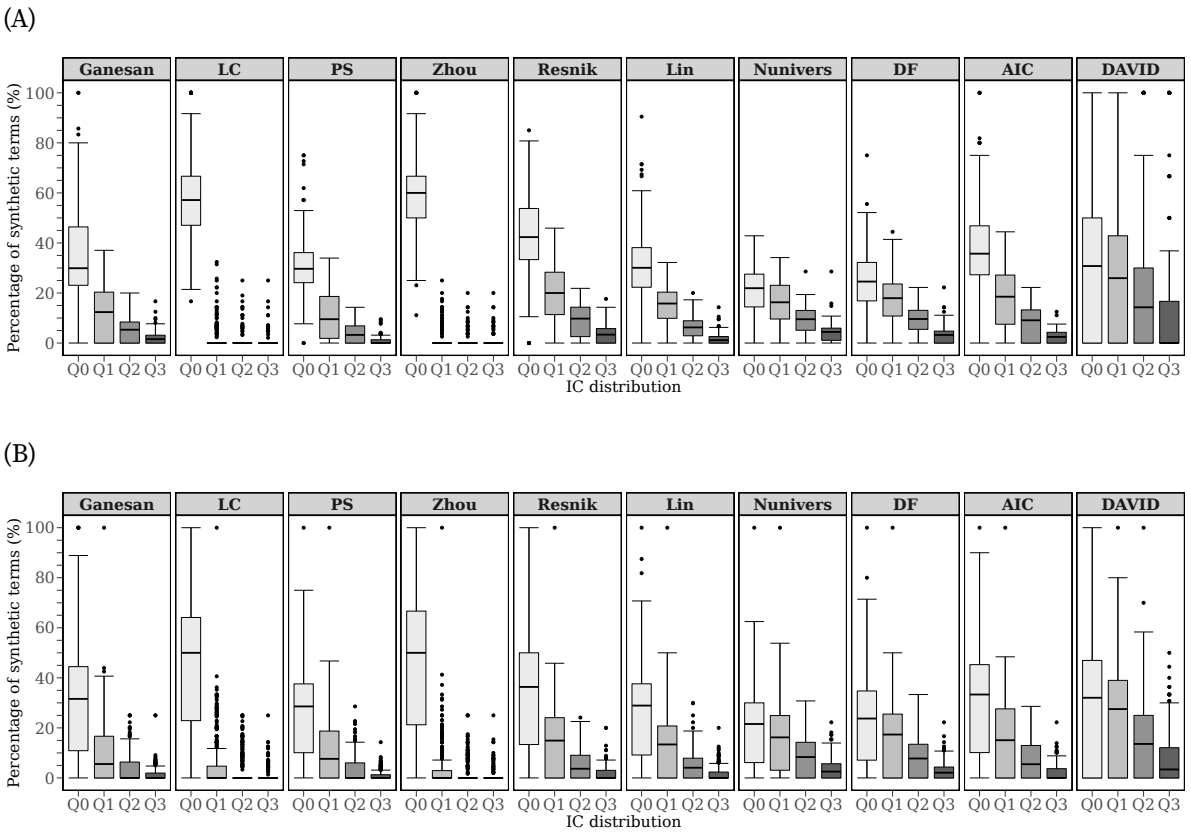


FIGURE 4.10: Percentage of synthetic terms using each semantic similarity measure and comparison with the DAVID enrichment tool for (A) [C-260] and (B) [B-346].

The box-plots shown in figure 4.10 display the ratio between the number of synthetic terms and the number of representative terms while figure 4.11 highlights the related gene coverage

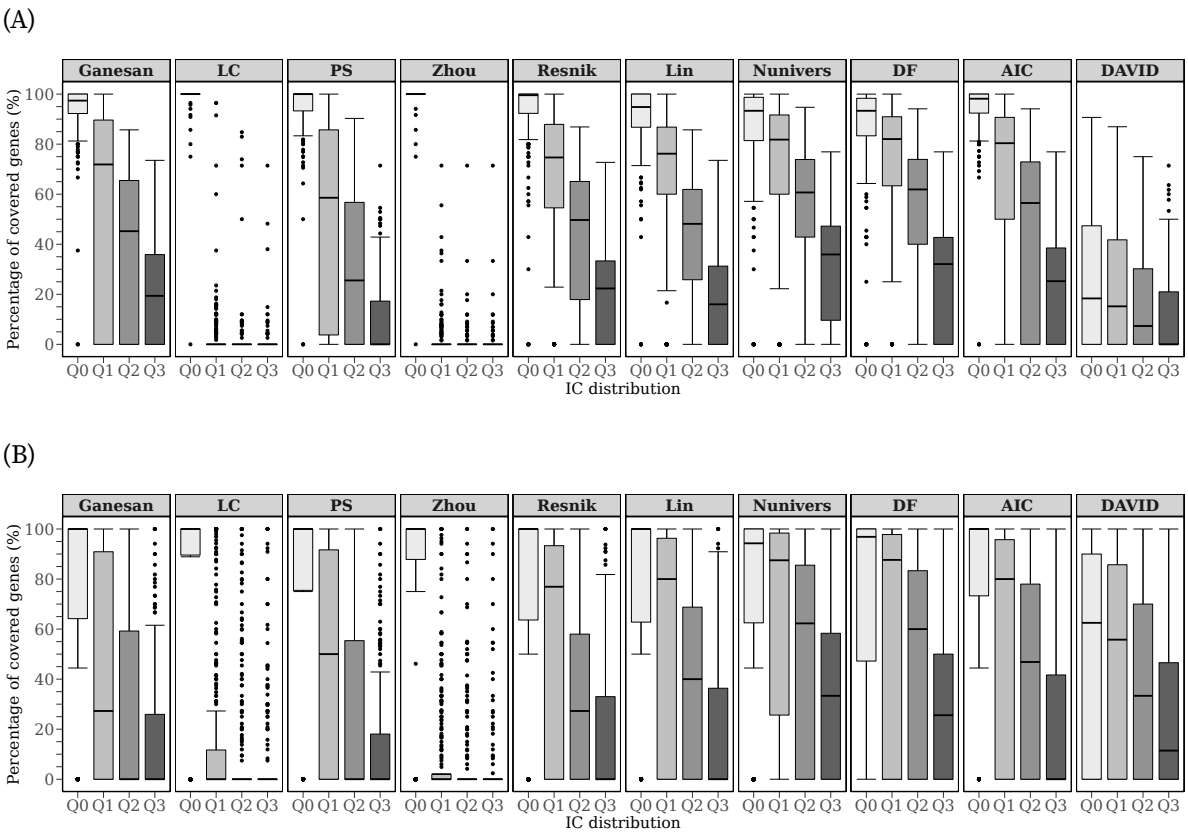


FIGURE 4.11: Percentage of covered genes using each semantic similarity measure and comparison with the DAVID enrichment tool for (A) [C-260] and (B) [B-346].

before and after applying the algorithm. For each measure, four box-plots were generated to qualitatively represent the percentage of synthetic terms and the percentage of gene coverage according to the four categories based on the continuous interval of the IC_{GOu} scores. The percentages of terms were aggregated from right to left since the cumulative effect allows for the representation of the first quartile Q_0, *i.e.*, the final percentage of the synthetic terms. As a reading guide, one may differentially consider the Q_0 box-plot in which the global annotation view was represented and the three other box-plots in which the fine-tuned analysis was presented. Based on the terms, the relative analysis aims to qualify the best measures when their median line is low and, in contrast, based on the genes, when their median line is high.

Focusing on figure 4.10, the downward trend observed in the first quartile Q_0 was less marked for the LC and Zhou measures. For these measures, the filtering stage had a less noticeable effect as a larger number of synthetic terms was retained. Of particular interest, their respective box-plots given for the three other quartiles showed that only a few terms with pertinent levels of information were retained. In contrast, Resnik, Nunivers, DF and AIC provided the best results with a larger number of synthetic terms having IC_{GOu} scores between the first and third quartiles regardless of the dataset. The other measures were more sensitive to the type of experimental data (*e.g.*, Ganesan showed interesting results for [C-260], whereas it performed poorly with [B-346]). Finally, the same analysis was performed using the results given by the DAVID tool [Jia+12]. Thus, we applied the enrichment analysis and only retained the resulting terms that annotate at least three genes. The decrease in terms from Q_0 to Q_3

is less pronounced than all investigated semantic similarity measures, corresponding to a higher number of more important terms with a high IC_{GOu} score.

Focusing on figure 4.11, the first observation concerned the Q_0 box-plots in which all semantic similarity measures nearly achieved a 100% coverage. However, in the other box-plots, the LC and Zhou measures showed singular box-plot profiles compared to the other measures. Indeed, most genes were related to synthetic terms with the weakest IC_{GOu} scores (*i.e.*, within the first quartile of values). Thus, these two measures positively retained a small number of terms (from Q_1 to Q_3) with the disadvantage of a low level of details. In contrast, Resnik, Lin, Nunivers, DF and AIC had better IC_{GOu} scores considering the percentages of the Q_1 to Q_3 categories of terms, regardless of the dataset. Among these measures, we focused on Nunivers and DF because most specific terms (within the last quartile of the IC_{GOu} score) still annotated more than 25% of the genes. From this perspective, the other measures appeared to be sensitive to the type of dataset. Finally, considering `DAVID`, we observed that its coverage of genes was lower than that of most semantic similarity measures. However, `DAVID` performed better with the [`B-346`], while simultaneously, it did not provide a gene coverage above 60% for half of the gene sets.

4.2.5 Discussion

An important finding of our study is that among the investigated semantic similarity measures, as shown in section 4.2.4, the *node-based* measures performed clearly better than the *edge-based* measures. This finding is consistent with Pesquita et al. [Pes+09] who observed that *edge-based* measures are not well adapted to compute the similarity among terms within GO. Indeed, GO terms situated at the same depth are not necessarily similarly specific and a given distance based on the count of edges between two terms does not attest the same semantic distance for another pair of GO terms. It can also be noted that LC is the *edge-based* measure that provides the worst results. Thus, the use of LCA by both Ganesan and PS appears to be effective in improving *edge-based* measures. The selected *hybrid* measure was not successful as shown in section 4.2.4, suggesting that the impact of the features in the *node-based* measures (*i.e.*, the IC_{S} and MICA for Zhou's measure) is negligible compared to the *edge-based* features (*i.e.*, the shortest distance between the two terms to be compared for Zhou's measure). Interestingly, we illustrated that *node-based* measures yield good results being similar although they make use of different features. Indeed, DF is singular as it only relies on the number of ancestors of the two terms to be compared while the other four measures are based on the *IC*. Finally, measures using *intrinsic* (Nunivers and AIC) versus *extrinsic* (Lin and Resnik) *ICs* did not reveal major differences between them. To confirm that *intrinsic* and *extrinsic ICs* could be used equally, it would be interesting to compute the Nunivers measure with the IC_{R} (because it is closer to Lin and Resnik than AIC) and determine whether similar results would be found or not.

4.3 The extension of the analysis framework

The framework defined in the previous section (section 4.2) was extended to develop the gene set annotation tool. As a guide, Figure 4.12 shows the differences between the method presented in section 4.2 and the one in this section. This original method, called *Gene Set Annotation* (GSAn), has been developed as an accessible web service tool (https://gsan.labri.fr) and implemented as an R package called RGSAn. This tool has been compared with some enrichment analysis tools in order to highlight its strengths in terms of annotation. Additionally, we evaluated GSAn within a case study related to the immune response.

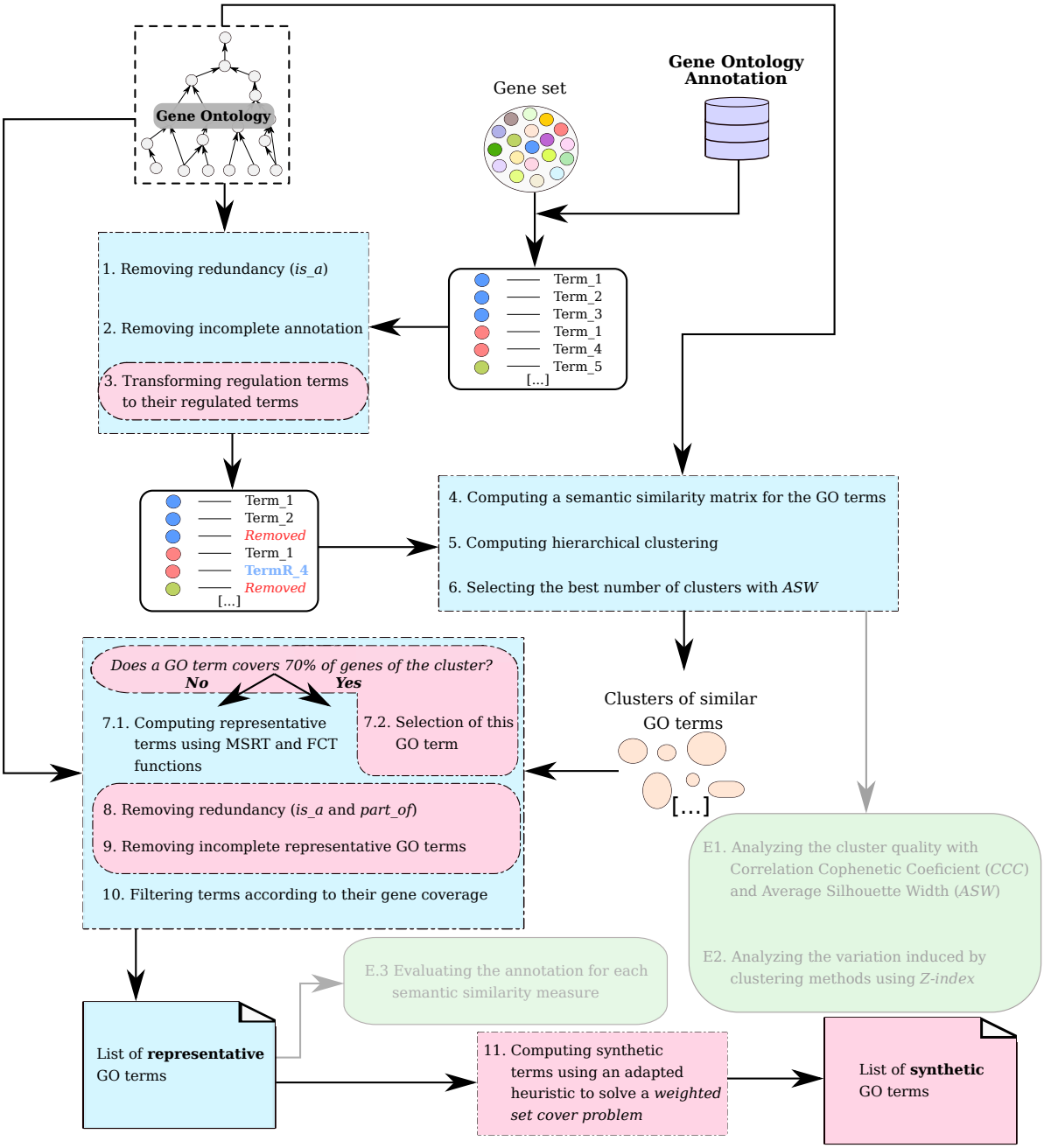


FIGURE 4.12: Comparison between the method proposed to annotate gene sets in section 4.2 and the method implemented in GSAn. The blue dotted rectangles correspond to the steps that are common to both methods, the pink dotted rectangles are the additional steps developed in GSAn and the dimmed green rectangles represent the steps used for the evaluation in section 4.2.

4.3.1 The GSAn method

GSAn is dedicated to the gene set annotation and is based on a method that makes use of the annotations from GOA [Cam+04] and the hierarchical structure of GO. The method is composed of five main steps which are described in the following paragraphs.

Elimination of the inappropriate annotations

The first step consists in removing inappropriate annotations and was computed as presented in section 4.2. The two differences are: (i) the way the distribution of the *IC* was calculated, and (ii) the conversion of GO terms related to biological regulation.

To obtain the same *IC* distribution for any organism, the distribution is thus calculated from all GO terms, instead of using the GO terms coming from the GOA of a particular organism. The main motivation to do that was the missing information provided by the annotation. Many organisms, especially the animal models, were strongly studied in their specific field, providing thus a high degree of GO annotation focused on their field [GD17]. For example, *Danio rerio* (usually known as zebrafish) is an animal model widely used for developmental biology and embryogenesis while the rat (*Rattus norvegicus*) used in toxicology. This missing information has an important impact on the distribution of *IC* since there are annotations with very specific GO terms while other annotations less studied have more general GO terms. Therefore, using the *IC* distribution from the GOA of a specific organism could then: (i) discard interesting GO terms, or (ii) include general GO terms. Finally, we decided to use all GO terms despite the substantial differences in GO annotation among the organisms.

We also considered an additional case of inappropriate annotation based on the regulatory relationships described within the GO ontology. Thus, we assumed that a gene associated with a term that regulates another term can also be associated with the corresponding regulated term. Thus, all regulation terms have been replaced by their regulated terms. For example, the term *regulation of ion transport* (GO:0043269) was replaced by *ion transport* (GO:0006811).

Clustering of terms according to semantic similarity measures

To compute the semantic similarity matrix of GO terms associated with a particular gene set, the following *node-based* semantic similarity measures were included within GSAn: Resnik [Res95] normalized according to the Jain and Bader's approach [JB10], Lin [Lin98], Nunivers [MM12a], DF [Que+15b] and AIC [Son+14]. Once the semantic similarity matrix was computed, it was used as input of the clustering method. For that, we applied the ALHM that exhibited the highest *CCC* compared with other clustering methods. At last, the best number of clusters was determined by using the *ASW* score [BGL11].

Identifying the most relevant representative terms

Considering that the number of representative terms may vary according to the size of the cluster, two distinct strategies were used to determine the best number of terms to be retained. First, if a single term inside a cluster annotated more than 70% of the genes, it was directly considered as representative. Secondly, if such a term did not exist, the *MSRT* and *FCT* algorithms described in section 4.2.4 were applied to identify an appropriate number of

representative terms for the cluster.

At the end of this stage that has been applied to each cluster, a new set of terms was obtained from the addition of representative terms of each cluster. Then, to retain the most relevant representative terms, we used two quality criteria: term redundancy and gene coverage.

1. *Removing inappropriate representative terms.* Some clusters of terms may have been generated from terms having a low similarity between them, resulting in very general representative terms. We thus removed the terms whose *IC* was lower than the first quartile of the *IC* distribution. Moreover, a new selection stage was then applied to eliminate potential redundancies. According to the type of hierarchical relationship (*is_a* or *part_of*), the removal of the ancestor terms may have a different impact on the number of annotated genes. To deal with this issue, a different strategy was applied according to the type of hierarchical relationship. For the *is_a* relationship, the representative terms being ancestors of other representative terms were removed. For the *part_of* relationship, only the parent or child terms annotating the largest number of genes were retained.

2. *Filtering representative terms according to the gene coverage.* To filter out the representative terms associated with a limited number of genes, we used a formula that depends on the size of the gene set provided as input. This filtering value computes the minimal number of genes for a given gene set and it gradually increases according to the number of genes. For a given gene set, the number of genes of this set is used to determine the minimal number of genes that must be annotated by each representative term. This threshold increases by steps based on the size of the gene set according to the following formula inspired by the equation (4.10):

$$f(gs) = floor(\sqrt{|\frac{Ngs}{10} - 1|}) + 2 \tag{4.11}$$

where gs is the gene set and Ngs is the number of genes in gs. Therefore, for a gene set containing from 2 to 19 genes, each representative term has to annotate at least 2 genes, from 20 to 49 genes, at least 3 genes, etc.

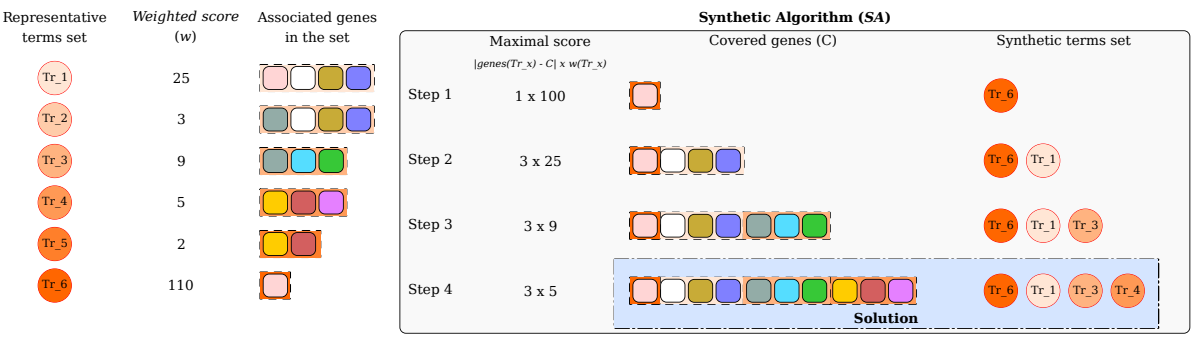


FIGURE 4.13: Example of *Synthetic Algorithm* (*SA*) with six representative terms. The *weighed score* is computed for each representative term according to equation (4.12). Then, the algorithm gets in multiple steps a subset of representative terms set (called synthetic terms set) that maximizes the *weighted scores* and covers all the genes annotated by the representative terms.

Identifying synthetic terms

At last, a final stage has been applied to the representative terms to get a more limited number of terms. In section 4.1.2, we carried out the reduction of annotation terms of a gene set while keeping the most relevant terms as **synthesis**. In section 4.2.4, we considered the set of representative terms associated with more than three genes as **synthetic terms**. However, the compromise between the specificity of a term and the gene coverage has to be taken into account. For that, we selected the terms that best summarize the biological information within the gene set by applying a heuristic algorithm based on the *Set Cover Problem* (*SCP*) [VLZ16] to the representative terms. The *SCP* is a NP-hard combinatorial optimization problem aiming, from a set of sets of elements, at deciphering the minimal set of sets covering all of the elements. Considering a term as a set of elements and an element as a gene, a set of representative terms is then a set of sets of genes. In this frame, we defined a solution of the *SCP* to identify the minimal set of terms covering the maximum number of genes.

For a set R of representative terms and a gene set G whose genes are annotated by at least a representative term, the synthetic terms were identified according to an iterative process. At each iteration, a score was computed for each representative term and the representative term with the biggest score was then added to a set S that gathers synthetic terms. This score is based on the number of genes annotated by a given representative term that are not yet covered by terms within S and on a *weighted score* associated with each representative term. This *weighted score* takes into account the *IC* of a term and the number of genes it annotates and was computed as follows:

$$w(t) = \frac{-log(\frac{annotated_genes_in_genome(t)}{nb_genes_in_genome})}{-log(\frac{annotated_genes_in_set(t)}{nb_genes_in_set})} \tag{4.12}$$

where $annotated_genes_in_set(t)$ (respectively $annotated_genes_in_genome(t)$) corresponds to the number of genes annotated by the term t in the gene set under investigation (respectively within the whole genome) and $nb_genes_in_set$ (respectively $nb_genes_in_genome$) is the total number of annotated genes within the gene set (respectively within the whole genome). In this formula, a relative measure (expressed as a ratio) has been used to evaluate the quantitative relation between two amounts of terms. The numerator actually corresponds to the *IC* proposed by Resnik [Res95].

At the end, the solution of our implementation of the *SCP* gives as results the minimal set of terms maximizing the sum of their weight and covering the gene retrieved by the representative gene sets. The pseudo-code of the *Synthetic Algorithm* (*SA*) with the customized *SCP* is presented in algorithm 3 whose final output is a list of synthetic terms (see an example in figure 4.13).

4.3.2 The `GSAn` web server

Based on the enriched methodology described in section 4.3.1, we developed a novel gene set annotation tool, called `GSAn`. In addition to be available on the web (https://gsan.labri.fr), `GSAn` provides interactive visualization facilities dedicated to the multi-scale analysis of gene

Algorithm 3: $SA(R,G)$

Input : R is a set of representative terms,
G is a set of genes covered by at least one term from R.
Used functions: $genes(t)$ is the list of genes in G annotated by the term t,
$w(t)$ is the weight score of the term t, as defined in equation (4.12)

1) Let S represent the set of synthetic terms and C
represent the set of genes covered by all terms in S.
Initialize $S := \emptyset$
Initialize $C := \emptyset$
2) While C is not the same as G :
a) Find the term $r \in R$ whose score is the biggest, in which score is defined by:
$score(r) := |genes(r) - C| \cdot w(r)$
b) Add the term with the biggest score to S and remove it from R
$S := S \cup \{r\}$
$R := R - \{r\}$
$C := \bigcup_{\forall s \in S} genes(s)$
return S

set annotations.

GSAn has been implemented in JAVA EE using the SpringBoot framework. From the client side, the web page exhibiting results has been implemented with the D3 [BOH11] and TreeColors[1] JavaScript libraries. The releases of GO and GOA are weekly updated and the JSON files created by GSAn are stored during 12 hours.

TABLE 4.2: List of organisms available in GSAn

Organism	File
Sus scrofa	goa_pig.gaf
Saccharomyces cerevisiae	goa_yeast.gaf
Rattus norvegicus	goa_rat.gaf
Mus musculus	goa_mouse.gaf
Homo sapiens	goa_human.gaf
Gallus gallus	goa_chicken.gaf
Escherichia coli	gene_association.ecocyc
Drosophila melanogaster	goa_fly.gaf
Danio rerio	goa_zebrafish.gaf
Canis lupus	goa_dog.gaf
Candida albicans	gene_association.cgd
Caenorhabditis elegans	goa_worm.gaf
Bos taurus	goa_cow.gaf
Arabidopsis thaliana	goa_arabidopsis.gaf

[1] https://github.com/e-/TreeColors.js/

GSAn server input

At first, users have to upload a gene or gene product list and to select the appropriate organism within the form. Fourteen organisms are currently stored in GSAn, downloaded from the GO web site[2] and from the European Bioinformatics Institute Web site[3], and listed in table 4.2. To be more flexible, users can also upload the annotation of any organism of interest using the GAF 2.1 format[4]. Users may choose any of the three GO sub-ontologies or any combination of them. If more than one sub-ontology is chosen, the analyses are computed separately and results are then merged. By default, GO annotations inferred automatically (evidence code: *Inferred from Electronic Annotation* or IEA) are included in the analysis but users may decide to exclude such annotations.

To customize the analysis, two advanced parameters are proposed to users: the gene support and the incomplete information filter. The gene support is the minimum number of genes that have to be associated to each representative term. The default value of this parameter is determined according to equation (4.11) (based on the size of the gene set) and can be modified. The incomplete information filter is used to discard terms presenting a low specificity in the ontology. Four levels of tolerance (none, low, medium and hard) can be applied, corresponding to the percentile values given by the *IC* distribution (1, 10, 25 and 50, respectively) of GO terms. As a result, terms below the chosen percentile value are discarded. Optionally, users can provide their email address to be notified when the analysis is finished.

The input parameters to be used in GSAn for the analysis are listed in table 4.3.

TABLE 4.3: Input parameters to be used in GSAn for the analysis

Parameter	**Description**	**Default value**
Gene list	A list of gene identifiers	-
Genome annotation	Organism name that will be used to recover the gene annotations. Fourteen organisms are proposed and any other organism can be uploaded using the GAF 2.1 format.	homo_sapiens
IEA	Boolean that indicates whether the IEA annotations should be used	true
Ontology	Sub-ontology of GO (BP, MF, CC) to be used for the analysis	BP
Semantic similarity measure	Measure to be used for computing the term similarity matrix	AIC
Advanced parameter		
Gene support	Minimal number of genes that have to be covered by any representative term	see equation (4.11)
Incomplete information filter	Tolerance degree to discard terms with a low specificity	medium
Email	Provided by users for being notified when the analysis is finished.	-

[2] http://www.geneontology.org/page/downloads
[3] https://www.ebi.ac.uk/GOA/downloads
[4] http://www.geneontology.org/page/go-annotation-file-gaf-format-21

(A) (B)

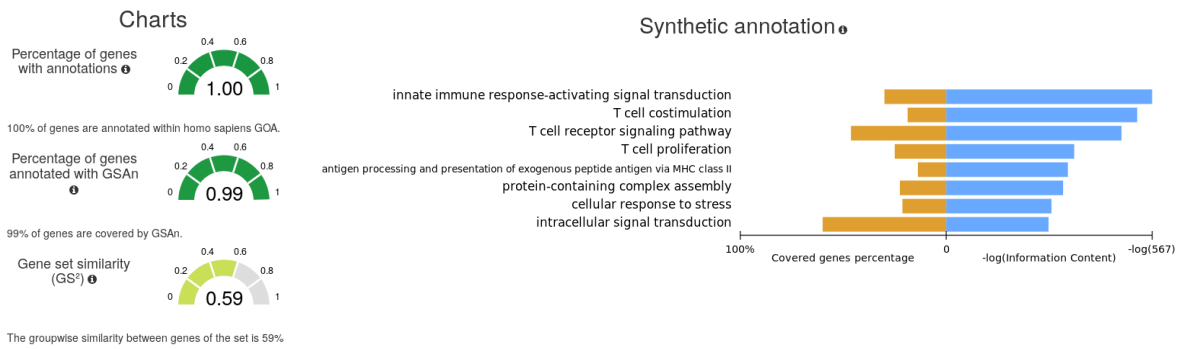


FIGURE 4.14: GSAn output results (1). (A) Three gauge plots show information about the annotated genes and the genes covered by GSAn as well as the groupwise similarity of genes in the set defined in Ruths et al. [RRN09]. (B) A diverging bar plot displays the *IC* and the gene coverage of each synthetic term.

GSAn retained 37 terms, 8 of them being synthetic
80 out of 81 genes are covered

Show 10 entries Search:

GOID	Name	Ontology	IC	Covered genes	Synthetic
GO:0002758	innate immune response-activating signal transduction	BP	566.68	24	true
GO:0031295	T cell costimulation	BP	357.68	15	true
GO:0050852	T cell receptor signaling pathway	BP	220.73	37	true
GO:0042098	T cell proliferation	BP	51.42	20	true
GO:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II	BP	42.33	11	true
GO:0065003	protein-containing complex assembly	BP	36.58	18	true
GO:0033554	cellular response to stress	BP	25.55	17	true
GO:0035556	intracellular signal transduction	BP	23.34	48	true
GO:0000187	activation of MAPK activity	BP	3740.61	7	false
GO:0006366	transcription by RNA polymerase II	BP	591.89	15	false

Showing 1 to 10 of 37 entries Previous 1 2 3 4 Next

FIGURE 4.15: GSAn output results (2). A table displays information about representative terms.

GSAn server output

GSAn results are presented according to multiple visual metaphors. At the top left, three gauge plots display the global gene set information (figure 4.14A). The first one indicates the percentage of genes which are annotated by GO terms while the second one provides the percentage of genes considered by GSAn. Finally, the gene set similarity consists in a groupwise measure, proposed by Ruths et al. [RRN09], that takes into account the gene annotation. A gene set similarity score of 1 means that all genes in the set have the same annotation and 0 means that terms have no common annotation. At the top right, a diverging bar plot displays the gene coverage and the *IC* score of each synthetic term (figure 4.14B). Information about the representative terms is provided in different formats within two separate pages: a table (figure 4.15) and the combined tree visualization MOTVIS, presented in section 3.2.4 (figure 4.16). The table summarizes the information of each representative term, being synthetic or not. MOTVIS aims to describe the hierarchical context of each representative term within GO. To obtain such a visualization, the GO structure (represented as a DAG) was converted into a tree according to the most informative parent term of each representative term (as described

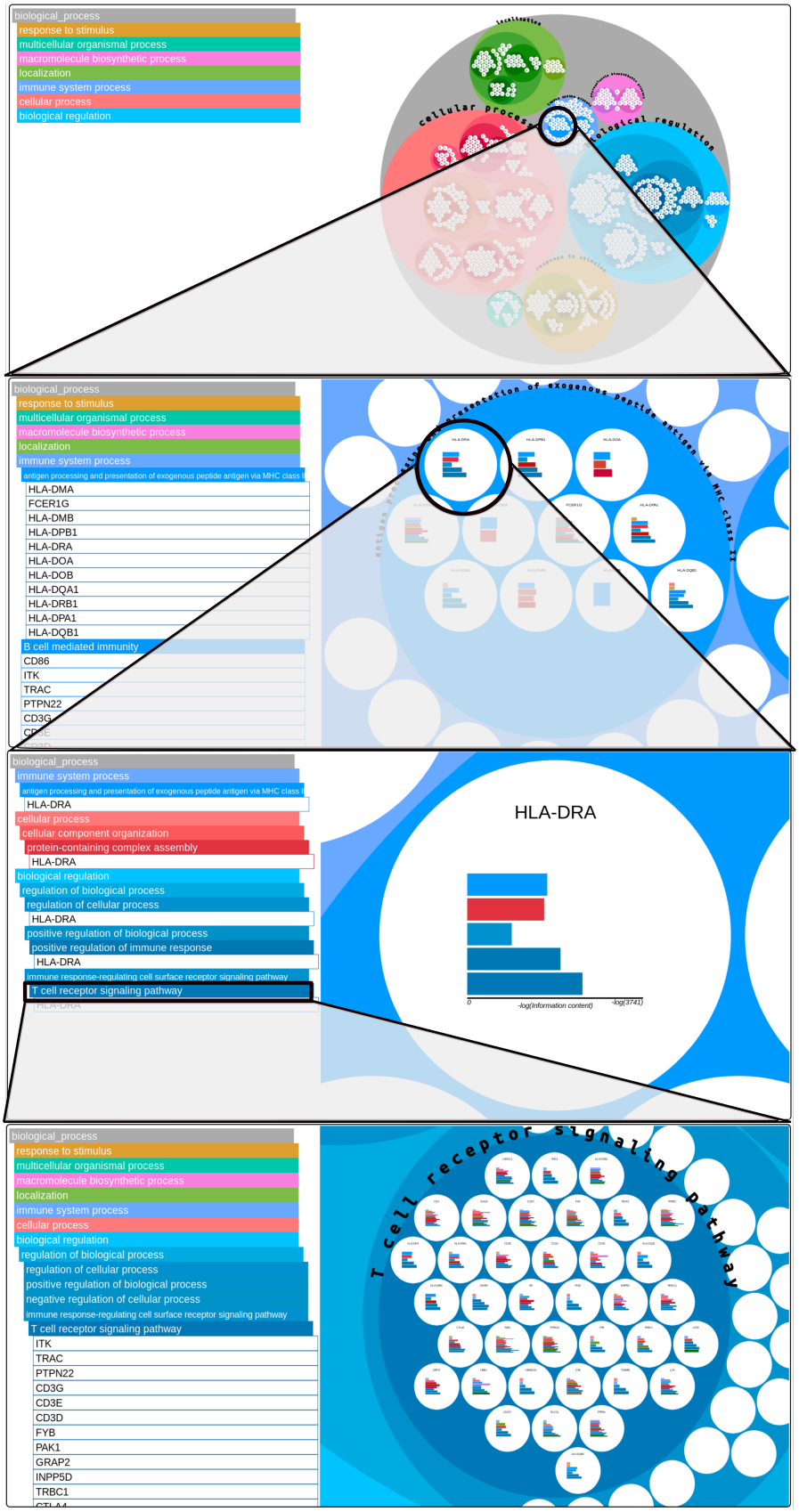


FIGURE 4.16: GSAn output results (3). An example of the combined visualization provided by MOTVIS illustrating the click and zoom interactions.

in section 3.2). Two types of tree visualizations are then combined: a collapsible indented tree and a circular treemap. White color forms represent the genes (instead of the gene set in chapter 3) inside their annotation terms. Thus, a given gene can appear inside several terms of different branches. Moreover, within each gene circle, a bar chart is displayed to represent its annotation terms (using their assigned colors). This visualization being interactive, it allows to explore annotation results thanks to interactions such as zooming within the circular treemap, or expanding the branch in the indented tree (figure 4.16). Additionally, users can download a JSON file and explore these results later on by uploading the file within the "Visualization" web page. Results can be downloaded as a CSV format.

4.3.3 Comparison of `GSAn` with enrichment analysis tools

We compared `GSAn` to the following enrichment analysis tools: `DAVID` [Den+03], `g:Profiler` [Rei+07], `clusterProfiler` [Yu+12] and `WebGestalt` [ZKS05] (table 4.4). As mentioned in the introduction, these tools include a reduction stage with the aim to reduce the number of terms by eliminating the redundancy, except for `DAVID`. This comparison investigates, among others, the impact of the reduction step to decreasing the number of annotation terms while maintaining the number of annotated genes.

TABLE 4.4: Comparison of gene set functional analysis tools.

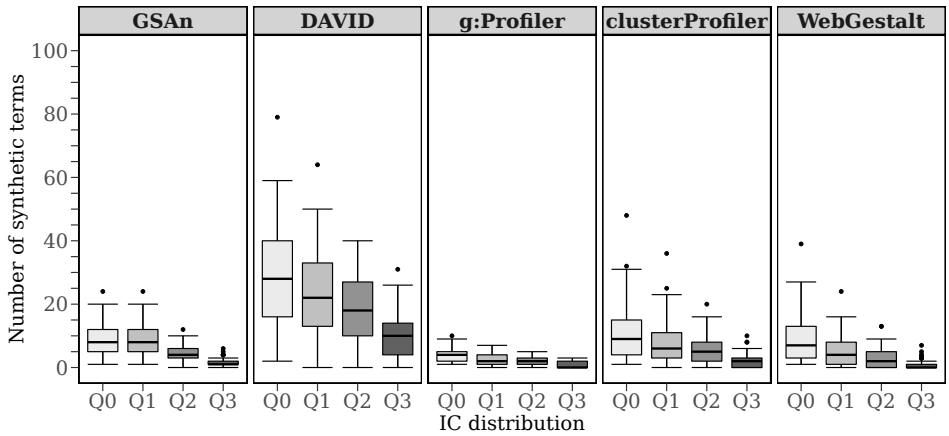
	`GSAn`	`DAVID`	`g:Profiler`	`clusterProfiler`	`WebGestalt`
Statistical method	-	Fisher's Exact	Hypergeometric	Hypergeometric	Hypergeometric
Genes id	Symbol	Many	Many	Many	Many
Accessibility	web wite, Rest Api	web site, Rest Api, R	web site, Rest Api, R	R	web site, R
Organism number	14*	hundreds	hundreds	2	12*
Updates	daily	-	-	-	-
Other resources	No	Yes	Yes	Yes	Yes
Reduction step	Synthetic algorithm, see section 4.3.2 for more details	-	Hierarchical filtering of the terms based on the p-value	Score filtering of the terms based on semantic similarity measures	Hierarchical filtering of the terms in the annotation file before computing the analysis

* The addition of any organism annotation file is possible.

To carry out this comparative analysis, we focused on the GO BP terms and retained only the gene sets involved in datasets [`C-260`] and [`B-346`] for which a gene set annotation was provided by all tools. Table 4.5 displays the number of gene sets annotated by each tool, as well as the number of gene sets that have been annotated by all of the tools. Thus, 62 and 226 gene sets were considered for [`C-260`] and [`B-346`], respectively. To analyze the gene coverage and the number of terms provided by each tool, we used the computed *IC* distribution to remove the incomplete annotations. We used four thresholds (from Q_0 to Q_3) corresponding to the quartiles of the *IC* distribution in order to filter out the results of each tool. As previously, for each threshold value (given by the limit of each quartile), the terms with a *IC* value below that threshold value are filtered out. Thus, Q_0 refers to the whole resulting GO terms and Q_1, Q_2 and Q_3 correspond to GO terms having an *IC* value over 18.4, 44.4 and 155.3, respectively.

Figure 4.17 and figure 4.18 display, for each tool, the number of terms and the gene coverage according to the *IC* distribution for datasets [`C-260`] and [`B-346`], respectively. Regarding the number of terms (figure 4.17), two classes of tools can be identified according to Q_0. First, `DAVID`, `clusterProfiler` and `WebGestalt` give a median number of terms ranging from 10 to

(A)



(B)

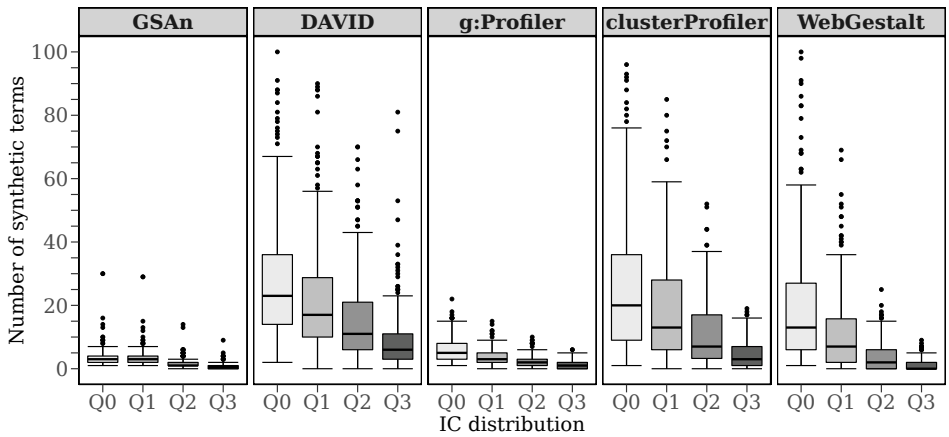


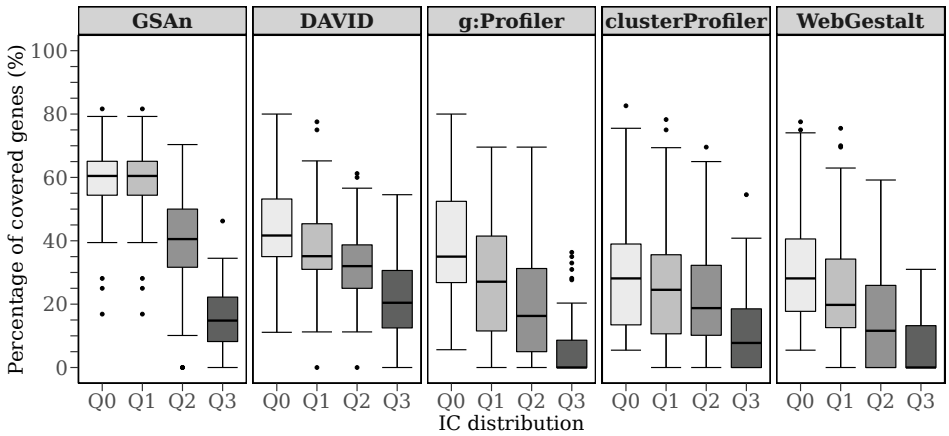
FIGURE 4.17: Box-plots providing the impact on the number of terms for each tool using (A) dataset [C-260] and (B) dataset [B-346] according to the quartile computed by the *IC* distribution in GO. Each quartile Q_x corresponds to the *IC* value according to which the terms are filtered out. Thus, Q_0 corresponds to the whole set of terms provided by the tools and Q_3 to terms with an *IC* value higher than the third quartile of the *IC* distribution in GO.

TABLE 4.5: Number of gene sets for which each tool provides an annotation.

	GSAn	DAVID	g:Profiler	clusterProfiler	WebGestalt	**Common**
# annotated gene sets of dataset [C-260] (260)	246	227	88	68	69	62
# annotated gene sets of dataset [B-346] (346)	337	336	284	233	270	226

25. The second group, involving GSAn and g:Profiler, has a smaller dispersion in the number of terms according to the high coverage of gene sets with a median value ranging from 0 to 5. This smaller number of terms combined with a high gene coverage is relevant because very few terms annotate almost the whole genes. Considering Q_1, all tools except for GSAn have a decrease in terms of gene coverage and number of terms. GSAn keeps the same values as for Q_0 due to the *IC* filter applied to remove the incomplete information (corresponding to the first quartile in the distribution).

(A)



(B)

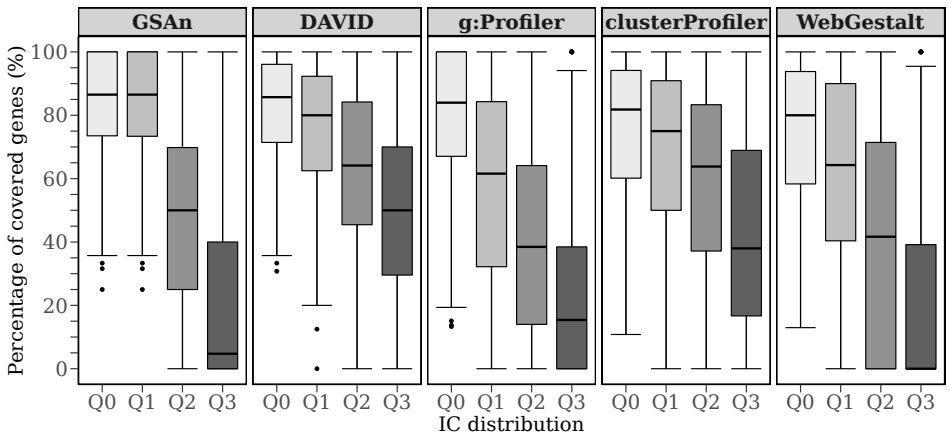
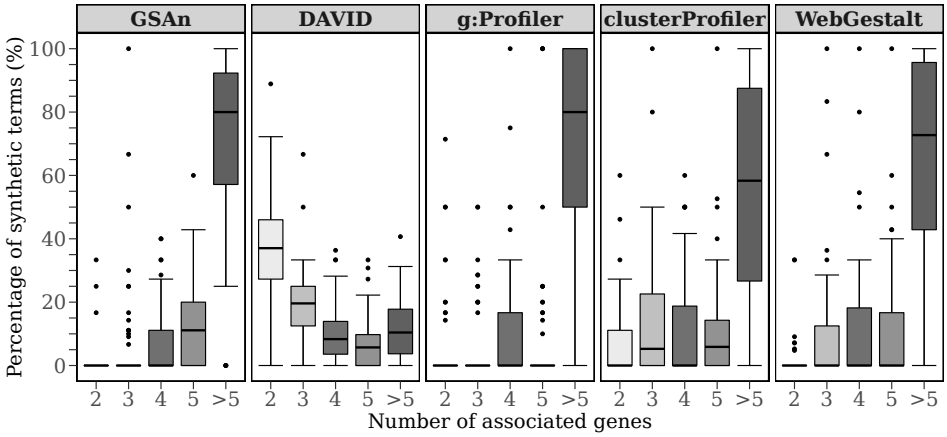


FIGURE 4.18: Box-plots providing the impact on the gene coverage for each tool using (A) dataset [C-260] and (B) dataset [B-346] according to the quartile computed by the *IC* distribution in GO.

On the other hand, regarding the distribution of the gene coverage in Q_0 (figure 4.18), GSAn recovers more genes than the rest of the tools for dataset [C-260], while no significant difference appear between tools for dataset [B-346]. For dataset [C-260], the median value is under 40% for all enrichment tools while GSAn covers over 60%. This might be due to two reasons; the presence of :(i) a high number of genes with missing annotations, and (ii) substantial differences between the genes of a given set in terms of annotation. This illustrates a main limitation of the enrichment tools, that tend to highlight the very well annotated genes at the detriment of other less studied genes. Regarding the number of terms, DAVID provides the highest number of terms with a median value over 30, while g:Profiler provides the lowest number of terms with a median value over 5. However, GSAn, clusterProfiler and WebGestalt have a median value under 10 term, which is an acceptable number of terms to be presented as result. For dataset [B-346], all tools present a median value higher than 80% of the complete gene set.

(A)



(B)

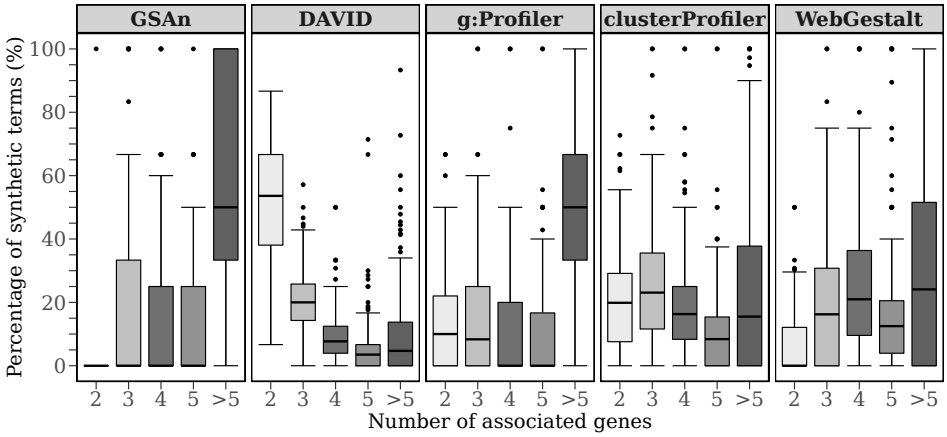


FIGURE 4.19: Box-plots showing the classification of terms by 2, 3, 4, 5 and more than 5 annotated genes for (A) dataset [C-260] and (B) dataset [B-346]. When terms are annotated mainly by 2 or 3 terms, it means that the result presents more specific terms. When terms are annotated mainly by more than 5 terms, it means that the terms are more general but that they are describing more genes.

At last, Q_2 and Q_3 present the results for the most specific terms in figure 4.17. In figure 4.18, the gene coverage is higher for DAVID and clusterProfiler with a gene coverage median over 40% for dataset [B-346]. They both lead to a better compromise regarding the number of terms and the gene coverage while maintaining relevant knowledge. For dataset [C-260], only DAVID presents the best compromise even if it has a gene coverage median over 20%. However, the high number of terms may suggest that each resulting term is likely to annotate few genes.

To further investigate this hypothesis, we analyzed the percentage of terms according to the number of genes that are annotated by these terms. Thus, figure 4.19A and figure 4.19B show the percentage of terms annotating 2, 3, 4, 5, and more than 5 genes for datasets [C-260] and [B-346], respectively. We observe that a median value of 40% of terms for dataset [C-260] and 50% of terms for dataset [B-346] are provided by DAVID, which thus annotate only two genes

and the rest of its median boxes do not exceed 20% for both datasets. On the contrary, the majority of terms provided by GSAn and g:Profiler annotate more than five genes. This implies that GSAn and g:Profiler provide terms with a lower specificity than DAVID but covering a larger number of genes for each term. Lastly, clusterProfiler and WebGestalt follow a similar behavior, with a predominance of terms annotating more than 5 genes in dataset [C-260] and a median value under 25% for each box for dataset [B-346].

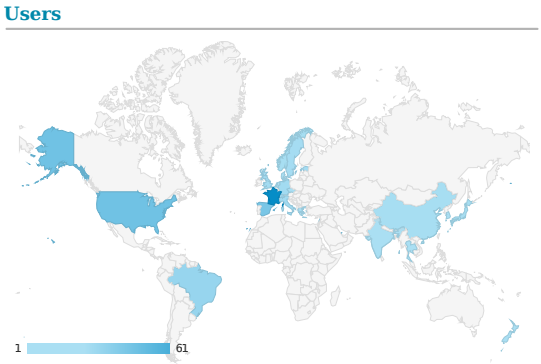
4.3.4 Case study

In this case study, we used GSAn to analyze a gene set of dataset [B-346] annotated by experts as *regulation of antigen presentation and immune response* [Li+13]. The *antigen presentation* is an important process to carry out an effective adaptive immune response. Cell types specialized to realize this process such as *macrophages, B cells* and *dendritic cells,* are processing the antigen into peptide fragments, then are boundering them into a *class II MHC molecule* and finally are displaying on their membrane [HAZ14; KL14]. Then, the *antigen-class II MHC molecule complex* is recognized and interacted by *T cells* (specifically *T cell helpers*). When these *T cells* are activated, they play important roles such as *B cell* antibody class switching or activation and growth of *cytotoxic T cells* [ALP14].

The used gene set contains 81 genes involved in the signal transduction in the immunological process against pathogens. The default parameters of GSAn were used and the chosen semantic similarity measure was Nunivers. GSAn retained 37 representative terms covering 80 out of 81 genes and 8 of them have been selected as synthetic terms (figure 4.15). The gauge plots (figure 4.14A) show a high gene coverage with all genes being annotated within the GOA file (first gauge) and 99% of genes annotated by GSAn (second gauge). At last, the third gauge displays a gene set similarity of 0.59, which means that genes share a quite high number of terms.

By focusing on the synthetic annotation displayed within the diverging bar plot (figure 4.14B), we can observe terms related to the proliferation and costimulation of T cells and the activation of signaling transduction by the innate immune response. These terms and the term *antigen processing and presentation of exogenous peptide antigen via MHC class II* (GO:0019886) are consistent with the manual annotation performed by experts and show that the annotation provided by GSAn is even more specific. Indeed, GSAn illustrates that the gene set is also involved in the proliferation of T cells. Moreover, more complete information can be observed from the representative terms through the information table (figure 4.15) or MOTVIS (figure 4.16). By exploring MOTVIS, we obtained additional information, such as terms sharing the same informative ancestor and the genes annotated by more than one term. For example, by focusing on the term *antigen processing and presentation of exogenous peptide antigen via MHC class II*, we can notice that eleven genes are annotated by this term (second screenshot). When clicking and developing in details each gene, we can see that six out of the eleven genes are annotated by *T cell receptor signaling pathway* (GO:0050852) and three of them by *T cell proliferation* (GO:0042098). Thus, with very few user interactions, we retrieve additional information about the biological role of some genes in the gene set.

(A)



(B)

Users and Sessions by Conuntry

Country	Users	Sessions
France	61	293
United States	22	38
Spain	19	65
Qatar	15	81
Brazil	8	10
Germany	4	44
United Kingdom	4	6
Italy	4	6
Estonia	3	4
Switzerland	3	3

FIGURE 4.20: Report on the number of GSAn's users in the world from November 18th 2018 (start date of the server) to September 29th 2019. (A) A *geomap* with the location of users of GSAn (a high blue density means a high number of users). (B) The top ten countries with the larger number of users and GSAn's sessions.

Application relevancy

GSAn is accessible since November 2018. To date, the number of users in GSAn is increasing, illustrating an interest to use this new approach solving some issues of enrichment analysis tools (presented in section 2.4). On September 29th 2019, GSAn has been used by 577 users, of which 420 reused it at least once. Figure 4.20A shows a *geomap* whose color represent the density of GSAn's users. Figure 4.20B shows the top ten of countries with the larger number of users and the total number of associated sessions.

4.3.5 GSAn in the R language or RGSAn

Additionally, an R package has been developed using the GSAn method. By combining natives functions of R and C++ (with the library rcpp), RGSAn allows to compute three main functions: (i) to import an ontology, (ii) to include the gene annotation of a specific organism, and (iii) to run the GSAn method for a given gene set. The ontology format used is the Open Biomedical Ontology (OBO) format. The choice of this format was the human readability of this format and an easier adaption in R. The annotation can be included in RGSAn in two ways:

- From the Go3AnnDbBimap R Class (this class can be found in libraries such as org.XX.eg[5]).
- From a gene association file (in the GAF 2.1 format).

Finally, the GSAn method can be executed after that the ontology and annotation have been loaded. The output can be obtained by exporting a tabulated table in CSV. Moreover, it is possible to represent the combined tree visualization described in section 4.3.2 by using the library htmltools.

[5]XX corresponds to the organism. For example, *Homo sapiens* annotation can be recovered from the org.Hs.eg. library

4.4 Conclusions

The main problems in finding gene signatures are related to the investigation of the biological function of gene sets. These problems can be solved using classical enrichment tools, such as `DAVID` or `g:Profiler`. However, these tools focus on the most studied genes which thus may provide annotations covering a limited number of annotated genes [BLG15; HTK18; Tom+18]. An additional problem inherent to this issue is the redundant annotation information because of which integration remains a largely manual process. To address these problems, bioinformatics offers various strategies ranging from enrichment analyses to semantic similarity measures. The latter approach has been intensively studied by the scientific community to provide a large range of measures. While these measures are often combined with enrichment methods, their *a priori* use may widely impact the interpretation of biological datasets. To investigate these challenges, we first developed a large-scale approach that uses semantic similarity measures within a robust interpretive analytic framework. We chose to use a straightforward set of nine measures covering various features and explored their pitfalls by examining criteria that may be good markers of information relevancy to domain experts. Thus, we analyzed the semantic similarity measures in terms of their capacity to synthesize information and provide the best trade-off for retaining detailed information.

Our main finding was that by using GO to annotate gene sets, better results were obtained with the *node-based* measures that use the terms' characteristics than with measures based on edges that link these terms. Moreover, by investigating more deeply the annotation of the gene sets provided by the *node-based* measures, the experiments did not detect any major differences, although these measures used different features. Notably, we did not consider some recent measures that use other relations than *is_a*. In particular, Wang et al. [Wan+07] proposed a semantic similarity measure that considers *part_of* relations. We believe that using all types of relations (*i.e.*, hierarchical and transversal) is an interesting approach and that axioms should also be considered, as described by Ferreira et al. [FHC13]. Axioms can be used to express the meaning of concepts and relations between concepts within ontologies [HSG15]. Thus, if the meaning of the GO terms was fully described (with a logical definition based on axioms), the GO terms could be better distinguished from their siblings (or other related terms). Some efforts have recently been made to enrich GO with such axioms [Mun+11; The15], opening up perspectives for proposing semantic similarity measures relying on their richness.

Then, we improved this approach and proposed a new web server as an alternative to classical enrichment analysis, called `GSAn`. Compared to enrichment analysis tools, `GSAn` has shown excellent results in terms of maximizing the gene coverage while minimizing the number of terms. `GSAn` has provided a gene set annotation that is more specific than the results given by experts (for a human gene set). Also, an originality of `GSAn` is to provide visualizations with interactive abilities to analyze the resulting gene set annotations. This visualization is based on the combined tree `MOTVIS` that provides zoom operations to browse terms and the genes they annotate according to the level of biological information that may interest users.

Finally, another interesting finding that emerged from the analysis of the human gene sets is that enrichment methods indeed mainly focus on the well-known subpart of genes. As

previously mentioned, the annotation of biological data is still an open question in scientific fields [HTK18; Tom+18]. Thus, we analyzed two large gene sets and specifically focused on human data, but the interpretation of biological experiments can change with the evolution of GO according to various organisms. Thus, the observations reported by Tomczak et al. [Tom+18] and Haynes et al. [HTK18] regarding the strong annotation bias in the GO annotations in which more than half of the annotations are related to approximately one fifth of the human genes have strongly guided the choice of the datasets. Consequently, methods that use *a priori* semantic similarity measures like `GSAn` improve gene set analyses by integrating evolving knowledge (including less-studied genes).

In this chapter, we made use of GO because it is widely used for understanding the biological roles of genes. However, other knowledge resources describing different types of information, such as diseases and pathways, could be included in order to enrich the annotation of a given gene set. Thus, the possibility of including knowledge resources within `GSAn` has been studied and evaluated, as is presented in chapter 5.

Chapter 5

Integration of additional knowledge resources within GSAn

The first collection of sequence data was realized in spring 1982 by the the European Molecular Biology Laboratory (EMBL), and later in the same year, GeneBank databases were established [Bur+85]. Nowadays, in the era of massive data, thousands of data sources, thesauri, ontologies and other knowledge resources are regularly created with the aim to make high amounts of biological knowledge available to the scientific community. For example, in 2018, the European Bioinformatic Institute (EBI)-EMBL reports that they store over 160 petabytes of biological information, including sequences, genes, variants, proteins, processes, diseases [Coo+18]. However, each of these knowledge resources partially covers the biological field. In section 2.3.4, we described the motivation for integrating different knowledge resources, the different standards developed and illustrated with projects that support such integration. Projects such as Bio2RDF or SyBioOnt aim to gather knowledge in order to give access to all the pertinent biological information trough complex queries. The INTEGRO initiative of Cinaglia et al. [CGV18] focuses on the disease-gene associations coming from various disease resources. Focused on phenotype information, Doğan [Doğ18] developed HPO2GO to propose a mapping between HPO and GO. Other existing solutions are based on methods implementing association rules to infer relevant relations between two different ontologies or two sub-parts of a single ontology [Far+12; BSS14; Aga+15].

This chapter presents an early stage of integration of two knowledge resources in order to provide additional information within the GSAn framework. One resource describes human diseases and the other one contains pathways. This knowledge resources have to be mapped to GO for enabling their integration within GSAn. At the same time, this integration should improve the coverage of annotated genes without significantly increasing the number of synthetic terms. This work has been evaluated by using the [C-260] and [B-346] human datasets presented in section 1.4. This chapter is structured as follows: section 5.1 presents existing solutions in enrichment analysis tools which make use of information from different knowledge resources for the annotation. Section 5.2 describes the resources we have chosen to include within GSAn and section 5.3 exposes the process implemented to integrate a new resource within GSAn. Section 5.5 presents the proposed approaches to map a new knowledge resource to GO. Section 5.6 shows an evaluation of the integration of resulting mappings in the GSAn analysis. At last, section 5.7 discusses the conclusions of this chapter.

5.1 Existing computational solutions that integrate knowledge resources

As previously presented in section 2.4, several challenges such as the different ways to describe the same information, the various origins of data and the lack of theoretical knowledge in the biological domain make the integration an arduous task. Considering the problems related to the volume of data, one has to also consider challenges given by computational methods. In particular, we noticed the lack of existing pipelines that enables to integrate even a few number of knowledge resources for enriching information associated with a given gene set.

GSAn (presented in details in section 4.3.2) aims to provide biological context for any gene set through GO annotation terms. However, the use of a unique knowledge resource may lead to incomplete information [HRM06; LG10] if one considers the richness of information given by the full set of biological and medical data. Therefore, the integration of other knowledge resources within the GSAn framework may surely improve the biological interpretation of gene sets. These challenges have also been raised by enrichment tools as some of them propose annotation terms recovered from several resources [Hua+07; Rei+07; Che+13; Ben+16]. However, the analyses performed by these tools apply statistical methods to find over-represented terms coming from various resources that are considered independently [Ben+16]. Then, in order to reduce the potential redundancies produced by equivalent terms coming from multiple knowledge resources, some tools have proposed solutions. For example, DAVID uses *Kappa statistics* in order to cluster annotation terms (from any resource) by using the gene occurrence [Hua+07]. Nevertheless, the clustering results (*i.e.*, groups of terms that are related according to the genes they co-annotate) are given to users as a list of clustered terms, without proposing a consensus term or considering the semantic relations between these terms to select the most pertinent. Then, users have to analyze a huge list of terms for each gene set they are interested in. GeneAnalytics™ annotates gene sets by using integrated knowledge resources within the GeneCards Suite [Ben+16]. For example, pathway information is stored and integrated in the knowledge resource PathCard that involves 3,215 human pathways from 12 resources into a set of 1,073 SuperPaths [Bel+15]. However, GeneAnalytics™ does not combine knowledge resources that describe very different biological information for a given gene set. For example, it is difficult to identify which pathway or biological process is involved within a disease of interest as no existing mapping provides this information.

A first attempt to integrate additional knowledge resources has been presented in section 3.2 with a pipeline that *a posteriori* integrates the enrichment results computed by using different knowledge resources. To do so, we adapted the lexical approach performed by OntoEnrich [Que+15a] to map new terms to GO terms and developed a new visualization metaphor, called MOTVIS, to explore the gene set annotations. Nevertheless, knowledge resources related to phenotypes provide information that cannot be mapped to GO using lexical approaches. For example, the DO term *platelet-type bleeding disorder 3* (DOID:0111056) is an inherited blood coagulation disease in which the GO term *platelet activation* is involved. By using OntoEnrich, no GO term (or only few GO terms that partially match) can be mapped to DO term because these resources describe different knowledge.

5.2 Description of knowledge resources to be included within GSAn

Two additional resources have been chosen with the objective to get new information regarding pathways and diseases within GSAn. Multiple knowledge resources have been developed in order to describe such information.

For describing diseases, knowledge resources such as *Medical Subject Headings* (MeSH) [Lip00], *National Cancer Institute thesaurus* (NCIt) [Gol+03], *Systematized Nomenclature Of Medicine–Clinical Terms* (SNOMED–CT) [Don06], *Online Mendelian Inheritance in Man* (OMIM) [Ham+05], *Orphanet Rare Disease Ontology* (ORDO) [Vas+14], *Human Phenotype Ontology* (HPO) [Rob+08] or DO [Sch+11] were good candidates to be included within GSAn because they are widely used. MeSH, SNOMED–CT and HPO were discarded since they do not specifically deal with genetic diseases while ORDO was not considered as it involves only rare diseases. Thus, we considered DO as the best candidate for the following reasons: (i) it covers a wide range of diseases, (ii) it provides annotations for genes and gene products, and (iii) it contains cross-references to other resources (thanks to the XREF metadata) such as SNOMED–CT, MeSH, HPO or ORDO. DO contains $9,384$ unique terms describing diseases associated to the human organism (using the v2019-07015 release).

For the pathways, *Kyoto Encyclopedia of Genes and Genomes* (KEGG) [KG00] is the broadest knowledge resource involving pathways with over $620,000$ pathways that are hierarchically related for more than $6,150$ genomes. However, KEGG has not been chosen as it is not freely available. Other knowledge resources representing pathways were thus considered, including *Reactome* [Jos+05], *WikiPathways* [Pic+08] and *Panther* [Tho+03]. We chose *Reactome* as this knowledge resource has the advantages to provide: (i) an easy web interface access, (ii) a large number of available data formats, and (iii) an alignment with GO[1]. *Reactome* contains more than $2,300$ unique pathways involving 16 species (see figure 5.1 for the statistical details of the organisms included in *Reactome*). Among the data provided by *Reactome*, we focused on the pathway sub-parts to analyze the interest of adding such information within GSAn.

5.3 Pre-process of resources before their integration within the GSAn framework

The addition of new knowledge resources within the GSAn framework requires to respect three constraints: (i) to provide complementary biological information to GO knowledge, (ii) to be organized as a graph structure, and (iii) to provide annotations relating terms with genes or gene products. Including *Reactome* and DO thus provide a complementary information to GO. However, new steps have to be implemented to adapt the graph structure or to integrate the annotation provided by these knowledge resources within GSAn. In order to facilitate the integration of these knowledge resources, the use of standard formats was considered.

[1] http://geneontology.org/docs/download-mappings/

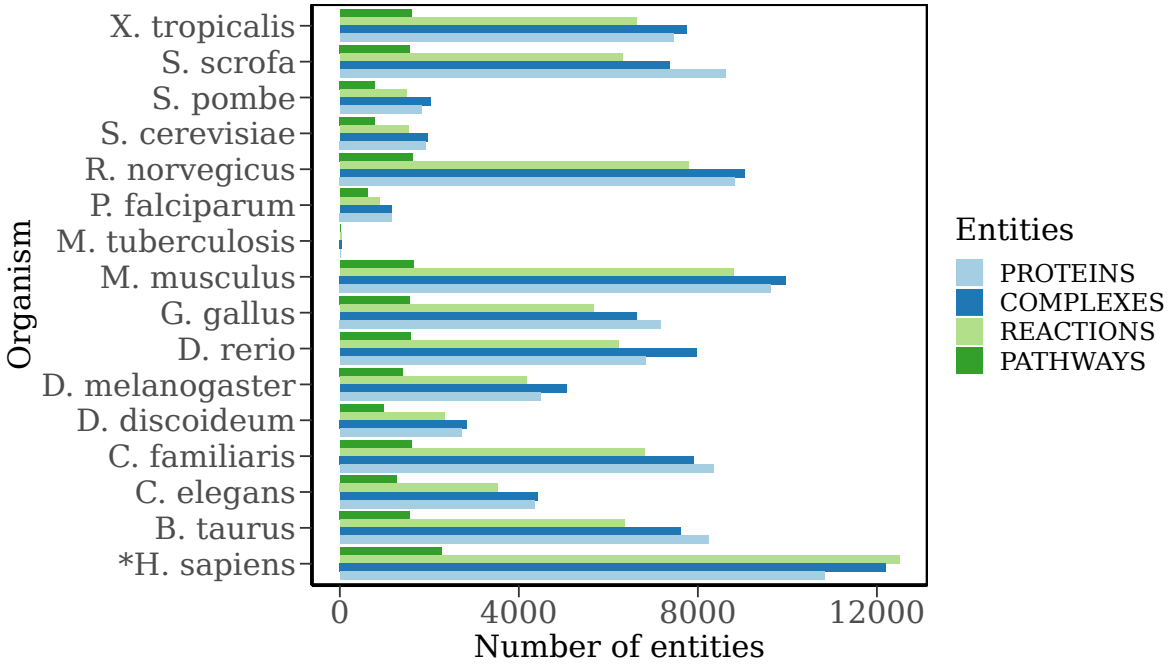


FIGURE 5.1: Data storage details of each organism within the *Reactome* release version 69. This barplot displays the number of involved proteins, complexes, reactions and pathways (Source: https://reactome.org/about/statistics).

5.3.1 Adapting the structure of *Disease Ontology* and *Reactome*

GSAn reads and manipulates the OWL format of GO (section 4.2). Since DO is also available in this format, its integration within GSAn was simplified. *Reactome* is described in different formats and one of them is BioPAX, defined in an OWL-based language. However, its integration within GSAn was not simple. The OWL format of GO describes GO terms as classes (OWL classes). In contrast, the BioPAX standard represents instance-based vocabularies whose entities are described as OWL individuals while other general entities are represented as OWL classes. In GO, the instances associated with the OWL classes would rather correspond to gene products. In order to overcome this issue, instead of creating an OWL file for *Reactome* to be included within GSAn, a simple plain-text format that included the pathway information and their hierarchical relation was used.

GSAn takes into account the different types of relations existing between GO terms. The *is_a* relation has been used at all GSAn stages (from the computation of semantic similarity between GO terms to the identification of representative GO terms for a given cluster). In contrast, the *part_of* and *regulates* (including *positively_regulates* and *negatively_regulates*) relations were only considered at specific stages in order to help filtering out redundancies. DO and *Reactome* use exclusively a single type of relationship to organize their terms. In DO, this relation is *is_a* and it defines, like GO, a taxonomy within the DO ontology. In *Reactome* that represents the graph structure in sub- and super-pathways, the structure is closer to a partonomy than to a taxonomy. Partonomies and taxonomies represent a hierarchy of their terms but they are differing with regard to the kind of relationship used to connect terms: respectively *part_of* and *is_a* [Lor+03; Zha+18]. While a taxonomy presents a hierarchy between

terms (describing the notion of specialization between terms), a partonomy is a decomposition of a term in different parts (describing the notion of terms being part of others) [Lor+03; Zha+18]

To facilitate the integration of *Reactome* within GSAn, we considered its partonomy structure as a taxonomy. This choice simplifies the adaptation of *Reactome*, and therefore avoids the need for developing additional steps within GSAn to explore a partonomy. This is a delicate choice since they do no represent the same information. However, the *part_of* relationship like the *is_a* relationship is transitive (*i.e.*, if a term A is part of a term B being itself part of a term C, the term A is part of the term C). Thus, it remains coherent to consider that a gene being associated with a *Reactome* term is also associated with all its ancestors following the *part_of* relations (according to the *true-path-rule*).

The characteristics of the graph structure of GO, DO and *Reactome*, *i.e.* the number of nodes and edges, the density (*#edges*/*#nodes*), the maximal depth and the *IC* distribution, are presented in table 5.1. To equally compare the three knowledge resources, we exclusively focused on the *is_a* relationship in GO. Moreover, due to the fact that *Reactome* presents organism-specific terms, only the *homo sapiens* sub-part of *Reactome* was considered. Considering that the number of terms in GO (and specifically of the *Biological Process* or BP sub-part) is larger than in DO and *Reactome*, which impacts in turns the depth and the *IC*, the *IC* distribution has been computed independently for each knowledge resource. This way, some terms coming from *Reactome* and DO, which would have potentially been removed (because of a small *IC* if computed according to GO), have been retained.

Knowledge resource	Nodes	Edges	Density	Maximal depth	*IC* distribution		
					1Q	2Q	3Q
GO	45,006	77,094	1.73	16	18.42	44.13	155.35
GO - BP	29,701	57,404	1.93	16	18.42	44.13	155.35
DO	9,079	11,702	1.29	11	13.01	20.17	29.77
Reactome	23,442	24,082	1.03	11	9.34	10.90	12.56
Reactome - human	2,222	2,271	1.02	11	9.34	10.90	12.56

TABLE 5.1: Description of the structure of the three knowledge resources included in GSAn.

5.3.2 Recovering annotations from DO and *Reactome*

Disease ontology

The associations between genes and disease terms are not officially provided by DO. However, they can be retrieved from external databases [Piñ+15; Ple+15] by using an algorithm implemented in INTEGRO [CGV18]. This algorithm exploits the cross-references to external knowledge resources provided by DO to recover the annotations. Using the INTEGRO algorithm, we thus generated gene-disease associations by merging information coming from the following knowledge resources: NCIt [Gol+03], OMIM [Ham+05], ORDO [Vas+14], DISEASES [Ple+15], and DisGeNET [Piñ+15]. In this way, we recovered the annotations within the basic (gene, DOID) pairs format.

Additionally, gene-disease associations were also generated by text-mining methods of DISEASES and DisGeNET databases. To avoid redundancies when text-mining approach is used, only associations that were not presented in the curated associations were kept.

- DISEASES generates associations between genes and diseases by mining MEDLINE abstracts using a *homemade* method [Paf+13; Ple+15]. For that, two dictionaries were first created: one including all disease names and synonyms of DO terms and another including the gene symbols from `STRING v9.1` [Szk+14]. Then, a tagging algorithm presented in [Paf+13] was used to extract terms from the dictionary within MEDLINE abstracts.

- DisGeNET generated associations between genes and diseases by mining MEDLINE abstracts using the `BeFree` system [Bra+14; Bra+15]. `BeFree` is a text-mining tool composed of two modules that decipher the information contained in biomedical documents. The first module is a *Biomedical Named Entity Recognition* (BioNER) module that detects occurrences of diseases and genes. This module is based on two dictionaries: a dictionary including the diseases available in *Unified Medical Language System* (UMLS) [McC89] and another one including the genes from *National Center for Biotechnology Information* (NCBI)-Gene [Mag+05], *Human genome organisation Gene Nomenclature Commite* (HGNC) [Pov+01] and UniProt [Apw+04]. Then, these dictionaries were used to establish associations between diseases and genes present in the dictionaries according to a set of MEDLINE articles. The second module is a relation extraction module based on morpho-syntactic information to identify relationships between the entities according to their co-occurrence in sentences to predict correct associations.

Reactome

The associations between genes and *Reactome* terms were acquired from two distinct resources: GO[2] and *Reactome*[3] websites. For the associations available on the *Reactome* website, the gene identifiers used by the association file (*i.e.*, `Entrez gene` identifiers) were converted to `official gene symbols`. Both sources of annotations were merged, thus resulting in a new gene association file in the GAF 2.1 format.

5.4 Evaluating the addition of a knowledge resource in GSAn without any integration

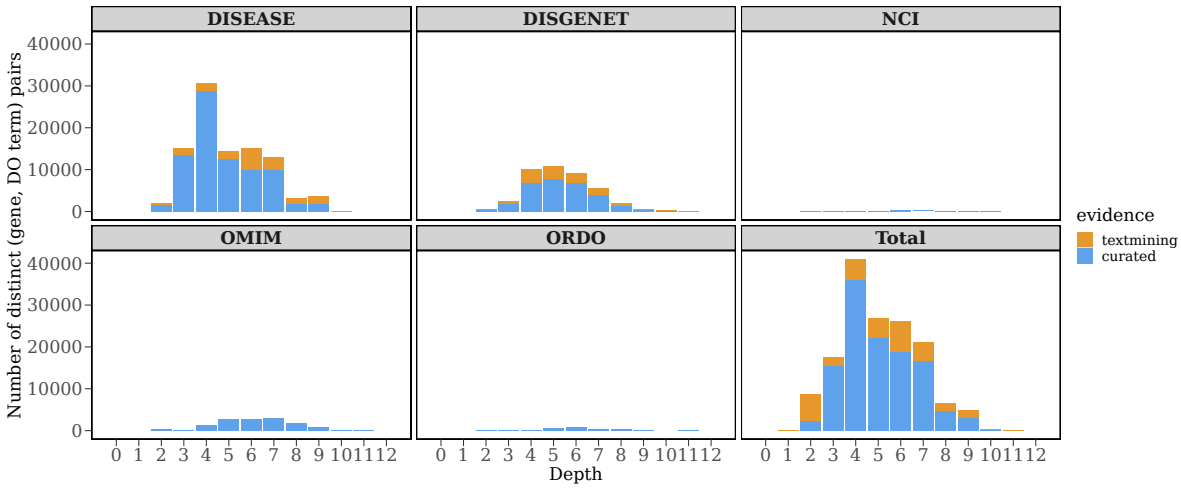
Once the new knowledge resources have been included within `GSAn`, we evaluated the impact of this addition for the annotation of gene sets. In this part, no integration has been carried out, meaning that mappings between GO and the two additional resources have not been used. In that case, we thus analyzed each knowledge resource independently and we merged the results after running the `GSAn` analysis.

Figure 5.2 shows the number of (gene-term) associations according to the depth of terms for each knowledge resource. Regarding (gene,disease) associations (figure 5.2A), we can observe that DISEASE and DisGeNET provide the highest number of gene-disease associations.

[2] http://current.geneontology.org/annotations/
[3] https://reactome.org/download-data

This figure shows that some disease resources do not provide many associations because of missing annotations (NCIt, OMIM and ORDO), or missing cross-references between DO and the other disease resources. As commented by Köhler et al. [KPL03], two different knowledge resources can present the same concepts even if no relation has been described between knowledge resources. For example, the `Alexander disease` term (a rare leukodystrophy) exists both in DO and ORDO but no `XREF` link has been proposed for this term between the two resources.

(A)



(B)

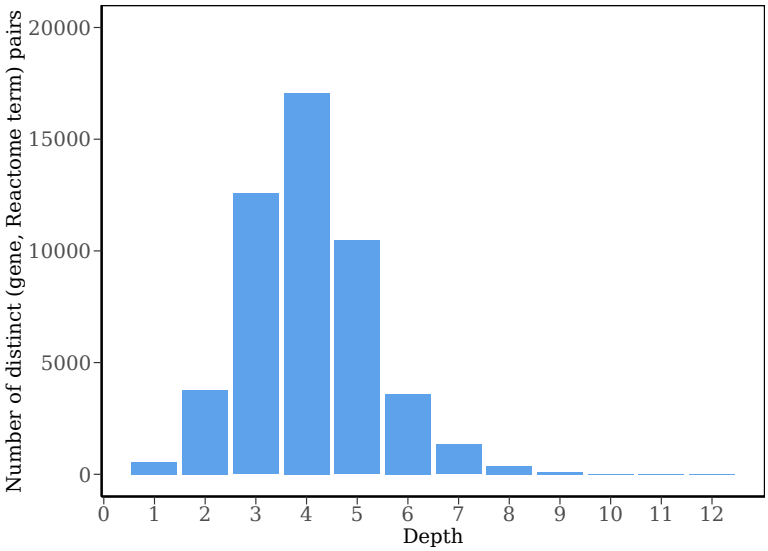


FIGURE 5.2: Number of (A) gene-DO term associations and (B) gene-*Reactome* term associations.

The terms showing the highest number of associations with genes have a relatively low depth, indicating that the corresponding genes are related to general terms. In DO (figure 5.2A) and *Reactome* (figure 5.2B), the largest number of associations involves terms with a depth of 4, corresponding to information that may be less detailed than other more specific terms (for more details, see appendix B):

- The DO term annotating the largest number of genes is *kidney cancer* (DOID:263, depth 5) with 1,285 genes. This term is the ancestor of 30 DO terms and a tiny number of them are related to 42 out of 1,285 genes, suggesting that most of the genes related to *kidney cancer* are not well known.
- In *Reactome*, one can find more heterogeneity as many terms with a depth of 3 or 4 have no descendant. For example, the *Reactome* term annotating the highest number of genes is *neutrophil degranulation* (R-HSA-6798695, depth 3) with 450 related genes and no descendant pathway.

To assess the impact induced by the addition of each knowledge resource, we evaluated: (i) the number of retained synthetic annotation terms, and (ii) the percentage of annotated genes (gene coverage) within GSAn. For that, we applied the same analysis based on *IC* distribution as used in section 4.3.2. Only the Q_1, Q_2 and Q_3 quartiles were considered (as Q_0 and Q_1 showed the same results using GSAn during the comparison analysis in section 4.3.3). Figures 5.3 and 5.4 illustrate the results for each knowledge resource and their combination (called DO+*Reactome*+GO) using the [C-260] and [B-346] datasets.

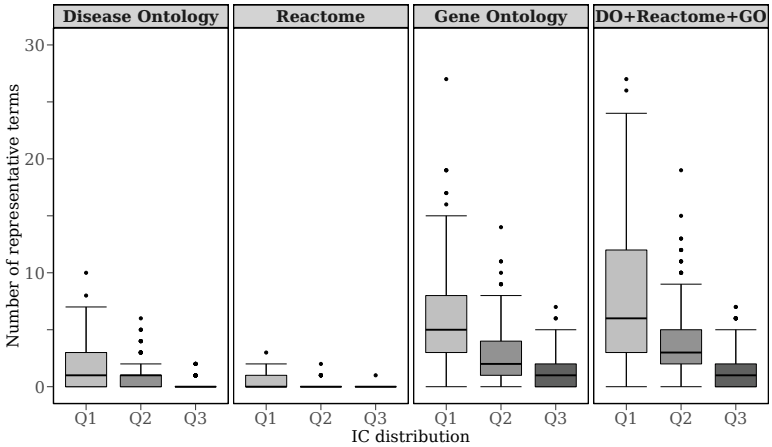
For both datasets, figure 5.3 presents the number of terms according to three cumulative categories of *IC*. These categories correspond to the Q_1, Q_2 to Q_3 quartiles of the cumulative results of *IC*, with an increasing level of information. Regarding the comparison between GO and DO+*Reactome*+GO, the addition of new knowledge resources increases the number of terms, generating more information to deal with. For example, in Q_2, the advantage of adding new knowledge resources within GSAn is significant with a *p-value* inferior to $2.2 \cdot 10^{-16}$ for the [C-260] and [B-346] datasets (using a Wilcoxon test[4]). This observation motivates the need for integrating these resources in order to eliminate potential redundancies.

A second analysis deals with the percentage of covered genes (figure 5.4) where the objective is to have the highest number of them. One can observe that the percentage of gene coverage is more important for DO than for *Reactome* within the Q_1 and Q_2 classes of results. The low percentage of gene coverage in DO and *Reactome* may be impacted by the specificity of terms in the annotation (as observed in figure 5.2 in which one can see that most of terms annotating genes are situated in a high position in the taxonomy).

Then, GO and DO+*Reactome*+GO have to be compared in order to investigate the benefit from adding other knowledge resources than GO. In both datasets, the percentage of annotated genes increases significantly in Q_2 with a *p-value* inferior to $2.2 \cdot 10^{-16}$ for [C-260] and [B-346]. Moreover, for the most specific category (corresponding to Q_3), the gene coverage is significantly impacted in [C-260] and [B-346] with a *p-value* of $2.586 \cdot 10^{-7}$ and $1.691 \cdot 10^{-13}$, respectively. These observations are motivating the need to include new terms coming from other knowledge resources because this allows to provide a richer and more various biological information from different contexts for a given gene set.

[4]**Wilcoxon test**: Non-parametrical statistic test that compares two distributions

(A)



(B)

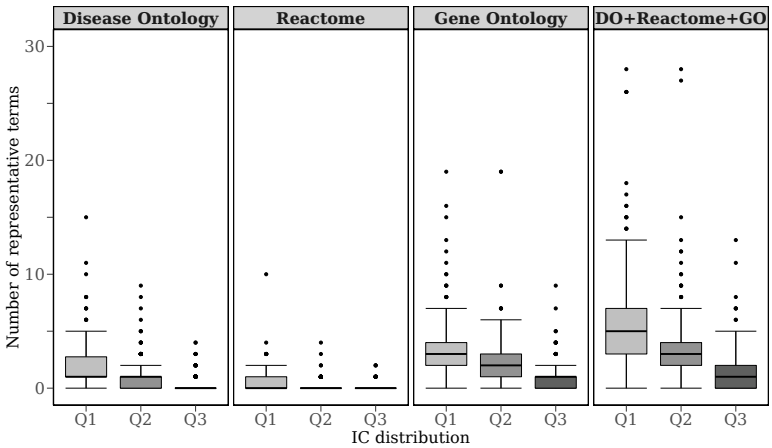


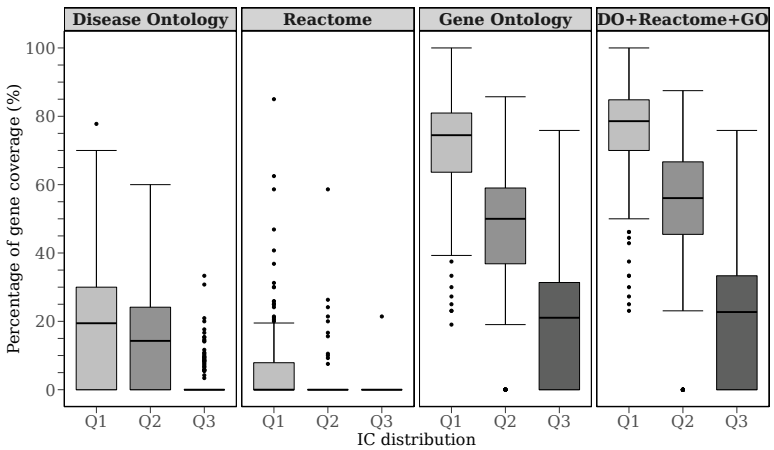
FIGURE 5.3: Number of representative terms computed by GSAn using DO, *Reactome* and GO independently as well as DO+*Reactome*+GO: (A) for the [C-260] dataset and (B) for the [B-346] dataset. The results are displayed according to the first three cumulative quartiles of the *IC* distribution (Q_1, Q_2 and Q_3).

5.5 Mapping methods to align new knowledge resources to GO

To reduce the number of terms to be included within GSAn, an integration of new knowledge resources through a mapping stage to GO may be useful. An alignment between two knowledge resources is defined by the collection of relations that connect terms considered as equivalent or hierarchically-related from the two resources (*i.e.*, *mappings*). An equivalence can be established: (i) between two terms, one of each knowledge resource, representing the same or very similar information, and (ii) between a term from a knowledge resource and a list of terms from another one.

Two strategies were considered to align a given knowledge resource to the BP sub-ontology of GO. The origin of the mappings is different according to the knowledge resource: the alignment between DO and GO needed to be computed while the alignment between *Reactome* and

(A)



(B)

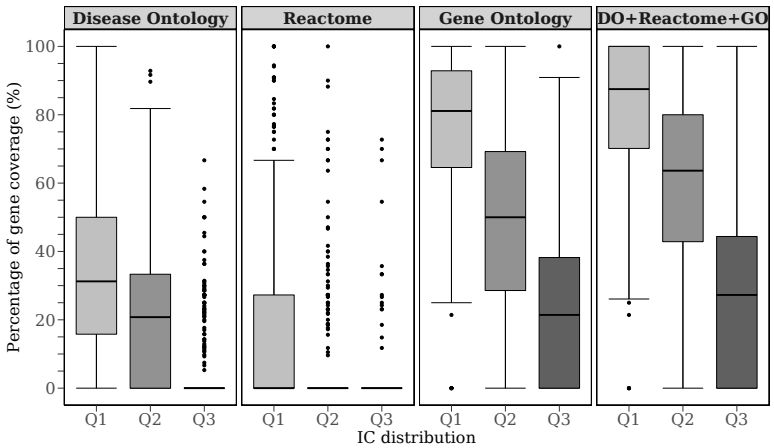


FIGURE 5.4: Gene coverage retrieved by GSAn using DO, *Reactome*, GO or DO+*Reactome*+GO: (A) for dataset [C-260] and (B) for dataset [B-346]. The results are displayed according to the first three cumulative quartiles of the term's *IC* distibution (Q_1, Q_2 and Q_3).

GO was already available from the GO web site.

5.5.1 Mapping Disease Ontology terms to Gene Ontology terms

To the best of our knowledge, no mappings exist between GO and DO terms. To address this issue, three methods have been developed and then compared. These mappings have been computed at the instance level, which means that they make use of genes annotated by terms from both knowledge resources. In other words, the mappings are related to the use of the terms by the genes, and may differ according to the organism.

1. A first method, called *MAP_DO_GO*_1, has been applied to each set of genes annotated by each DO term. It exploits the (gene,disease) associations created in section 5.3.2 and applies an enrichment analysis considering the genes of a particular DO term as a set. To do so, we used `g:Profiler` that performs the hypergeometric distribution to

calculate the GO terms that are statistically over-represented within the gene set. The `g:Profiler` hierarchical filter `STRONG` was then applied as post-treatment to reduce the redundancy and number of terms [Rei+07; Rei+16].

2. A second method, *MAP_DO_GO_2*, has also been applied to each set of genes annotated by each DO term. It follows the same principle, but uses `GSAn` with the semantic similarity measure *NUnivers* instead of running `g:Profiler`.

3. The last method, *MAP_DO_GO_3*, creates a mapping between the occurrences of GO and DO terms in the genome by making use of the (gene,disease) associations of each knowledge resource. To do so, we implemented a preliminary stage to infer association rules, based on the *frequent item set* (`FIS`) algorithm [Nau+13]. A `FIS` is a set of items that frequently appears when a data model is considered. `FIS` mining is often applied in *market-basket* data models considering a transaction of clients that purchase different market items [Nau+13]. Thus, the framework of `FIS` was applied to our gene set annotation problem where the genes are transactions and each associated term is an an item. Then, the itemset is a list of terms that are considered frequent if the support (*i.e.* the number of related genes) is higher than a given threshold. To compute the frequent itemsets, the `FIS` algorithm chosen was `FPGrowth` [Nau+13].

For each method, a filtering stage has been applied to the DO terms to discard the ones that annotate less than five genes. In this way, to analyze sets of genes, the filter helps to avoid the mappings that may be created because of missing information in the used annotation files. Table 5.2 displays the number of DO terms that have been mapped to a GO term (as a minimum) according to the three mapping methods.

A first observation relies on the advantage of using methods including the hierarchical structure of knowledge resources (such as `GSAn` and `g:Profiler`) to compute the DO to GO mappings. Also, even if similar results were obtained by `GSAn` and `g:Profiler`, the highest number of mappings was recovered by `GSAn`, which may suggest that using the hierarchy in an *a priori* stage is more useful than in an *a posteriori* stage.

Mapping method	**Number of mapped DO terms**
`Mapping with` *MAP_DO_GO_1*: `g:Profiler`	1,522
`Mapping with` *MAP_DO_GO_2*: `GSAn`	1,618
`Mapping with` *MAP_DO_GO_3*: `FIS`	175

TABLE 5.2: Mapping of DO and GO terms

5.5.2 Mapping *Reactome* terms to GO terms

The mapping between GO and *Reactome* can be recovered from the GO website[5]. It is also possible to extract additional mappings directly from the GOA files. Indeed, some annotations involve a gene and a GO term, as well as a *Reactome* term. In such cases, a mapping between the corresponding GO and *Reactome* terms can thus be extracted. Then, both mappings were merged to benefit from all information.

[5] http://current.geneontology.org/ontology/external2go/

Table 5.3 shows the number of *Reactome* terms that have been mapped to GO terms. Considering the number of *Reactome* terms in the *homo sapiens* organism (2,222 terms presented in table 5.1), 38% of them have been mapped with GO after merging both mappings.

Mapping extracted from	**Number of mapped *Reactome* terms**
GO mapping file	420
GOA file	734
Total number of mappings after merging	845

TABLE 5.3: Mapping of *Reactome* and GO terms

5.6 Comparison between an *a priori* and *a posteriori* integration

Once mappings have been obtained between knowledge resources, an integration stage has been carried out. Two strategies have been applied: (i) an *a priori* integration, and (ii) an *a posteriori* integration.

- The *a priori* strategy makes use of the mapping between GO terms and *Reactome* or DO terms before applying the GSAn method.
- The *a posteriori* strategy makes use of the mapping after running the GSAn method. During this process, GSAn separately exploits different terms from the resources and the mapping is used at the end of the analysis.

In this preliminary study, we realized this comparison by examining the number of mappings between GO terms and *Reactome* or DO terms that have actually been used in the different strategies. This comparison allows us to see why the integration is crucial in GSAn when additional knowledge resources are included to then provide richer and non redundant information by making use of correspondences existing between the different resources.

5.6.1 Integrating DO within GSAn

The mappings between DO terms and GO terms were obtained from three methods based on the gene sets associated with DO terms (section 5.5.1). As very few mappings were recovered by FIS algorithm, only the mappings provided by *MAP_DO_GO*_1 and *MAP_DO_GO*_2 mappings were finally considered. Thus, a single DO term is associated with a list of GO terms and some GO terms are also associated with several DO terms. The existence of these *many-to-many* associations can render the integration stage more difficult as a match between a single GO term and a single DO term is not enough to consider two terms as equivalent. An illustration of this issue is presented in figure 5.5. In this example, a set of five genes are annotated by the following three disease terms and four biological processes:

- Diseases:
  - *pulmonary embolism* (DOID:9477),
  - *cholestasis* (DOID:13580),
  - *adult respiratory distress syndrome* (DOID:11394).

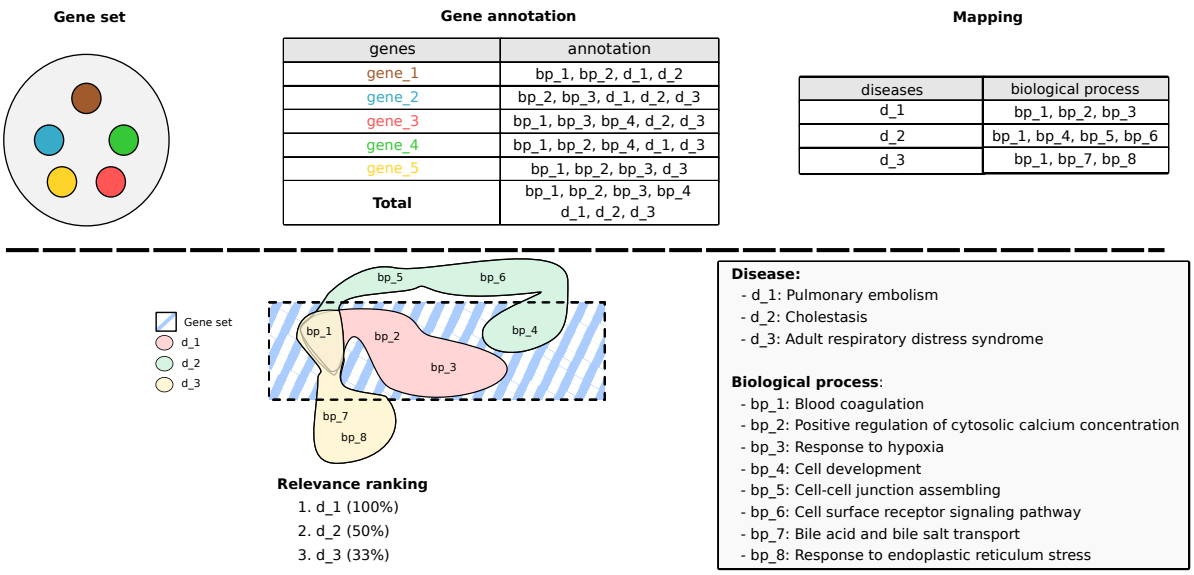


FIGURE 5.5: Example of annotation performed by GSAn when using DO and GO. The DO term d_1 (*pulmonary embolism*) is considered to be the most interesting disease since all its mapped biological processes annotate the studied gene set.

- Biological processes:
  - *blood coagulation* (GO:0007596),
  - *positive regulation of cytosolic calcium concentration* (GO:0099588),
  - *response to hypoxia* (GO:0001666),
  - *cell development* (GO:0048468).

We can observe in figure 5.5 that the GO term *blood coagulation* has been mapped to the three DO terms. As we did not evaluate the quality of the mappings created to link DO and GO, we decided to consider only DO terms mapped to a high number of GO terms associated with the gene set. To do so, we computed a mapping percentage score for each DO term, defined as follows:

$$percentage(DO_x) = \frac{|\text{GO terms associated to } DO_x \text{ in the gene set}|}{|\text{GO terms associated to } DO_x|} \cdot 100 \tag{5.1}$$

For a given DO term (DO_x), the numerator involves the number of GO terms mapped to this DO term that annotate the gene set and the denominator is the number of GO terms mapped to the DO term. Thus, the DO terms can be ranked by the percentage of mapped GO terms annotating the gene set. In the example of figure 5.5, the DO term *pulmonary embolism* is the most relevant since 100% of the GO terms to which it is mapped are part of the annotation of the gene set.

To analyze the impact of an *a priori* versus *a posteriori* strategy, the results are analyzed according to five categories corresponding to the five ranges of the previous percentage: $]0,30[$, $[30,45[$, $[45,60[$, $[60,75[$ and $[75,100]$. Thus, in figure 5.5, the DO term *pulmonary embolism* is included in the $[75,100]$ category and the DO terms *cholestasis* and *adult respiratory distress syndrome* are included in $[30,45[$ and $[45,60[$, respectively.

Figure 5.6 shows the number of DO to GO mappings obtained by the *MAP_DO_GO_*1 and *MAP_DO_GO_*2 methods using the *a priori* and *a posteriori* strategies. As expected, the *a priori* strategy includes a higher number of mappings since, at this stage, there is a larger number of terms associated with genes in the set. It could be interesting to also apply a *mapping weight* to each DO term before the analysis, in order to rank the DO term(s) according to their relevance. Thus, for a DO term, this *mapping weight* could be determined considering the number of its GO mapped terms and the number of genes it annotates in the used gene set. Comparing the *MAP_DO_GO_*1 (based on `g:Profiler`) and *MAP_DO_GO_*2 (based on `GSAn`) methods to generate a mapping between a DO term and several GO terms, the mapping provided by the *MAP_DO_GO_*2 method performs better than the *MAP_DO_GO_*1 method (as shown in figure 5.6). The upward trend observed in the first range $]0,30[$ indicates a large number of non relevant potential mappings between GO and DO terms. This abundance tends to decrease as the range is bigger. Focusing on the last two ranges, the range $[75,100]$ is slightly higher and involves less outliers, suggesting a good compromise between a low tolerance in the mapping coverage and an acceptable relevance given by DO terms. It seems that a disease term with a high mapping percentage has a strong association with the gene set, which may also mean that if this mapping has a low percentage, it is possible that the disease is present only by chance.

5.6.2 Integrating *Reactome* within `GSAn`

To focus on the pathway sub-part of *Reactome*, we extracted the mappings of GO terms with *Reactome* terms as explained in section 5.5.2. The mappings are one-to-one associations, meaning that a GO term is associated with a *Reactome* term and vice-versa. To evaluate the number of mappings found between GO and *Reactome* terms that annotate a gene set with `GSAn`, we searched for these mappings in two different ways:

1. **Direct mapping**: it corresponds to cases where a direct mapping exists between a GO term and a *Reactome* term that both annotate at least one gene of the studied gene set,

2. **Indirect mapping** : if the GO term mapped to a *Reactome* term is not in the pool of annotation terms, an indirect mapping has been established between the *Reactome* term and an ancestor of the mapped GO term that is part of the annotation terms and has the highest IC value.

Due to the fact that the mapping between *Reactome* and GO is one-to-one, when a direct mapping exists between a *Reactome* term and a GO term, it is not necessary to search for an indirect mapping for this *Reactome* term. As shown in figure 5.7, the indirect mappings provide minimal mapping coverage compared to direct mappings. This finding suggests that the direct mapping is sufficient to map *Reactome* and GO terms.

Another finding of this analysis, similarly as observed about the mapping between DO and GO (section 5.6.1), is that the *a priori* strategy finds the greatest number of mappings between GO and *Reactome*. Nevertheless, the strategy to integrate this resource in `GSAn` is different to the one proposed for DO and GO. The reasons for this difference are twofold: (i) the mappings being provided by GO, their quality is established, and (ii) the mappings used one-to-one.

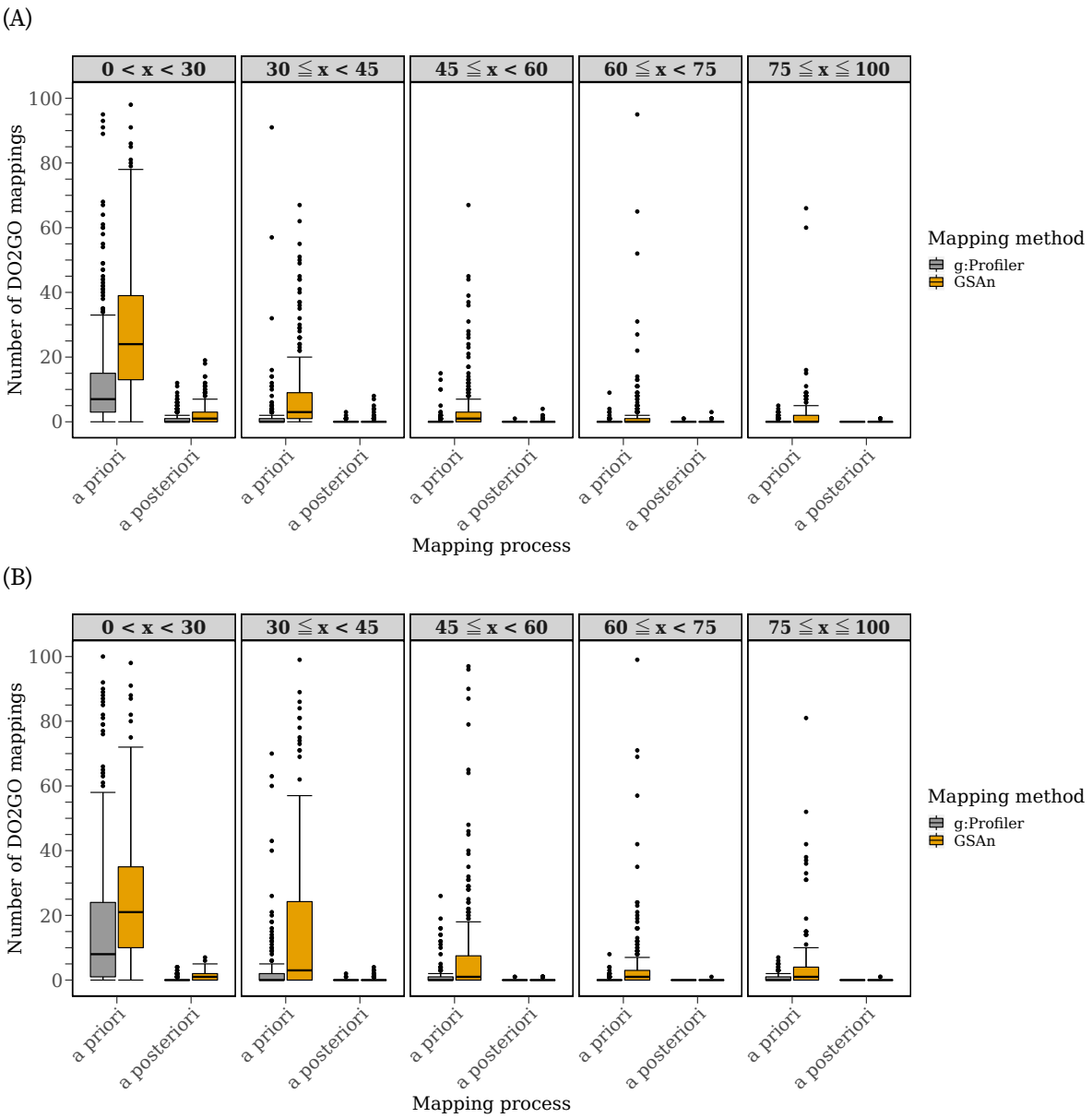


FIGURE 5.6: Box-plots of DO to GO mappings that were acquired by the *MAP_DO_GO*_1 and *MAP_DO_-GO*_2 methods using an *a priori* strategy and an *a posteriori* strategy within the `GSAn` method: (A) for dataset [`C-260`], and (B) for dataset [`B-346`]. As a DO term may be mapped to more than one GO term, the analysis was performed according to five categories depending on the percentage of GO terms that annotate the gene set and are associated with the mapped DO term.

5.7 Conclusion

We presented in this chapter a preliminary work that has been carried out to integrate new knowledge resources within `GSAn`. In a first time, we evaluated the impact of such integration to improve the biological information that can be extracted from a gene set. In a second time, we evaluated the impact of mappings between knowledge resources to get richer annotation results. Two knowledge resources have been studied, DO and *Reactome*, in order to provide additional biological context regarding disease and metabolic information, respectively. For DO, we had to establish an alignment to GO while for *Reactome*, mappings were available

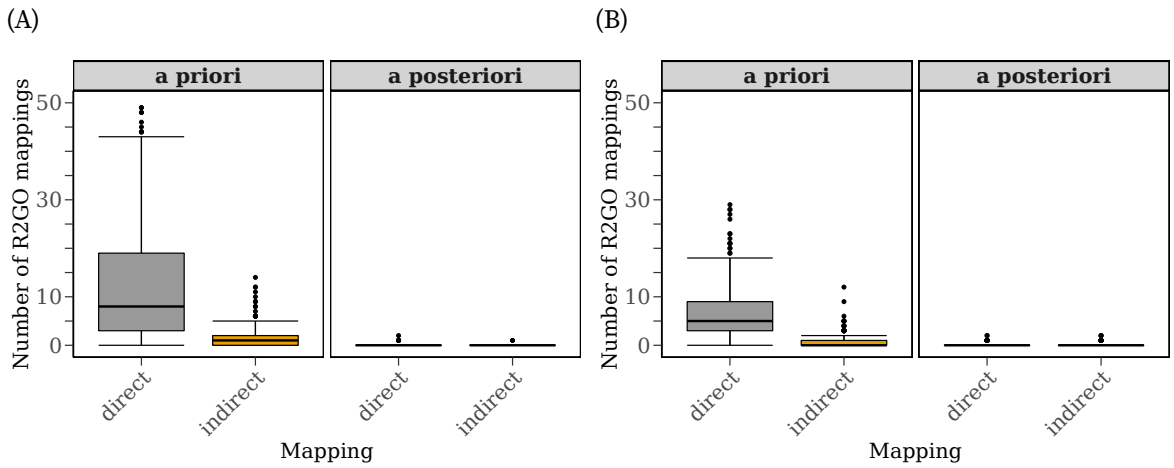


FIGURE 5.7: Box-plots of mappings between *Reactome* and GO terms used before and after applying the GSAn method: (A) for dataset [C-260], and (B) for dataset [B-346]. Two types of mappings were used: direct and indirect.

on online databases. In both cases, we showed that their integration within the GSAn framework increased the gene coverage. Nevertheless, using separately the knowledge resources increases the number of annotation terms and this could make the results more difficult to interpret.

To solve this issue, two integration strategies have been studied: an *a priori* and *a posteriori* use of mappings besides the GSAn computation. This comparison may help to decide which strategy is more relevant for future works. Thus, the *a priori* mapping would require to add or to adapt steps into the GSAn method for taking into account the fact that annotations described by multiple resources (or related by different resources) may be more relevant. In contrast, the *a posteriori* mapping would require a simple step to merge the annotations described by multiple resources. Regarding the results, the *a priori* integration is more relevant as more mappings are computed between a knowledge resource and GO for a given gene set. Thus, the *a priori* integration has the advantage to provide a greater mapping coverage between two knowledge resources.
Towards an integrative analysis, the mapping stage is crucial to merge two or more knowledge resources. Many other knowledge resources that have a direct relationship with genes may increase the biological relevance of any gene set like drugs, reactions, cell types or specific phenotypes [BRA05; Wis+07; Dav+08; Rob+08; SE09; Mat+10; Die+16].

Chapter 6

Conclusions and research perspectives

The use of gene sets in biological conditions has been very popular in the last two decades. This interest is due to the emergence of high-throughput technologies (*i.e.*, microarray or `RNAseq`) that generate a large amount of data to be processed in a single analysis. However, the quantity of information associated with a gene in the form of annotation is a major challenge when working with a gene set. The goal of this thesis was to investigate, by exploring different fields of study, how to improve the annotation, and then, the interpretation of gene sets. In this final chapter, we recall the initial research questions motivating our work and recollect the findings which allowed us to address them. We then conclude by outlining some existing opportunities and research perspectives.

6.1 Review and findings

In section 2.3, we described a synthesis of four fields of study to help to better understand the implication of a gene set in a biological context. These fields are the following:

1. Annotating gene sets with an **enrichment analysis** that presents over-represented annotation terms.
2. Proposing an alternative to enrichment with **ontological solutions** to group genes according to the **semantic similarity** of their annotations.
3. Representing the annotation of a gene set into **visual metaphors** to facilitate the exploration of results for better understanding its interpretation according to a given biological context.
4. Implementing **techniques** for integrating multiple knowledge resources in order to enrich the information available with the aim to interpret the biological function of gene sets.

Then, in section 2.4, we presented the different challenges of these objectives that have guided our work and motivated choices made during this thesis. The remainder of the manuscript describes how we dealt with these challenges through three main works.

The first work addressed two main questions associated with a visualization problem, that is, how to observe the impact on the results of gene set annotation when: (i) visualizing one or multiple gene sets, and (ii) including structured resources such as ontologies in the visualization. In the first case, by using multiple gene sets, the visualization makes it possible to observe which are the gene sets are involved in the same phenotype. Including the structure

between annotation terms was useful to better understand the results and to organize the annotation terms while reducing redundancies between related terms. Additionally, as stated by Kerren et al. [Ker+17], the integration of different visual metaphors being still occasionally, we developed an integrative visualization, called MOTVIS, that combines two different visual metaphors: a circular treemap and an indented tree. MOTVIS, which has been used to represent the results of annotation from multiple gene sets, has the advantage to solve the limitations presented in each visual metaphor when used individually. This advantage illustrates the interest of using different visual metaphors to facilitate the comprehension of biological results when complex data are represented.

The second work focused on the challenges posed by two distinct approaches when annotating genes: the enrichment analysis and the ontological solutions, based on semantic similarity. Although enrichment analysis has the advantage to facilitate the functional understanding of gene sets, it may be criticized for two important reasons: (i) the lost of poorly annotated genes, and (ii) a high level of redundancy in the results. Ontological solutions based on semantic similarity have been intensively studied by the scientific community to provide a large range of measures. While these measures are often combined with enrichment methods, their *a priori* use may widely impact the interpretation of biological datasets. Thus, we chose to develop a multi-step method that computes a synthetic annotation for a gene set combining semantic similarity approaches with techniques of data mining and heuristic algorithms. First, due to the diversity of semantic similarity measures, we evaluated the different approaches grouped in three categories: *node-based*, *edge-based* and *hybrid*. The evaluation of each type exhibited that *node-based* measures provided better results with no significant distinctions between them. Besides, comparing these methods with the DAVID enrichment tool showed that our workflow with *node-based* semantic similarity measures gave better annotation results. Then, this method was adapted to create a new online tool, GSAn, dedicated to compute a synthetic annotation for a given gene set. In comparison with classical enrichment tools, GSAn offers three main advantages: (i) to recover a maximal number of genes, (ii) to minimize the number of annotation terms, and (iii) to provide a good compromise in terms of relevant information within annotation and the involved genes. GSAn also offers visualization facilities, including MOTVIS adapted to visualize a single gene set, enabling a simple interpretation of results and the opportunity of using interactive tools for exploring these results.

The last work presented in this manuscript is an extension of the second work and addressed the last challenge involving the integration of other knowledge resources within GSAn. The main motivation of this work was that only GO was used to recover annotation terms in the GSAn framework. As described in section 2.2.1, GO represents the biological processes, functions and localizations of a gene or gene product. However, many knowledge resources describe other facets related to genes. Consequently, it was important to analyze the possibility to include information provided by other resources than GO that may improve the understanding of gene set annotation. Nevertheless, the knowledge resources cannot be included *per se*. As the integration of different resources may increase the amount of information associated with a gene set, it may also imply a harder interpretation of results. Therefore, in an early stage, we studied integration strategies to include two knowledge resources within GSAn, namely *Reactome* and *Disease Ontology* (DO), and to find mappings between terms of

these resources and GO terms. We applied two strategies to find an alignement (either extracted from existing databases, or computed in this work) between each new resource and GO. We stated that a mapping process used before applying the GSAn method gave a higher number of inter-relations between knowledge resources.

In the course of this work, GSAn was developed as an alternative tool to annotate gene sets. This tool uses a different approach to statistical analysis and may be very useful in genomics analysis. Even though GSAn is already used by the community[1] and produces pertinent annotation outcomes, some limitations and perspectives have nonetheless to be taken into consideration.

6.2 Research perspectives

Despite the promising results obtained in these works, they also raise issues needing improvements. Before concluding this thesis, we present the perspectives that could solve some limitations or improve the current methods or visualizations developed during the past three years.

Rendering a more informative visualization of the gene set annotation

In chapter 3, in order to facilitate the interpretation of gene set annotation, we developed MOTVIS to visualize the annotation of multiple gene sets. This integrative visualization was then used in chapter 4 for a single gene set as output of the GSAn online tool. This visualization used a simplification of the GO structure to represent the annotation terms, which allows to reduce the complexity of the GO structure for an easier exploration of the results. Keeping the most informative parent for a particular term was a solution to convert the DAG of GO into a tree. Nevertheless, simplifying the GO structure could be problematic given the loss of context it implies for terms. Indeed, sibling terms may be represented in different places in the visualization (when another parent term is more informative) whereas their meanings are probably close.

Therefore, it might be interesting to explore alternative metaphors adapted to the visualization of a DAG to be implemented in MOTVIS. Moreover, it could be interesting to include in the visualization additional information such as the genes involved in a gene set or different types of relations existing between terms (*e.g.*, *part_of* and *regulates* relations in GO). Finally, MOTVIS may benefit from additional functionalities, such as interactive thresholds to filter terms by their properties (*e.g.*, the *IC* value) or a search engine enabling to search for a particular term, gene or gene set. Even so, these implementations would make it possible to find a compromise between data completeness and user understanding.

Developing semantic similarities considering different properties of terms

In section 4.2, we restricted our evaluation to nine semantic similarity measures while the number of semantic similarity measures is extensive (see [Har+15] or the supplementary data

[1] As of 29 September 2019, GSAn has been used by 577 users, 420 of them return more than once on the web site.

of [MCM17]). A future perspective is to investigate the impact of a wider spectrum of semantic similarity measures and to find the best way to combine them for obtaining the highest similarity between terms. In this sense, Shin et al. [Shi+15] developed an *hybrid* semantic similarity measure, called *consensus similarity*, for short text clustering. This similarity measure combines four pairwise similarities: distance between terms, cosine similarity, pair relatedness and surface similarity. Martinez-Gil [Mar16] developed `CoTo` (Consensus Trade-off), a tool that computes a fuzzy membership to detect differences between semantic similarity measures and then aggregate some of them having different features. The tool stated better results comparing to classical semantic similarity measures by using different human rating datasets.

Moreover, in section 2.3.2, we showed the difference between semantic similarity and semantic relatedness. While semantic similarity only considers the taxonomy to compute the closeness between two terms, the semantic relatedness can involve a broader range of relations (*i.e.*, involving hierarchical and associative relations) [PPM04; Pes+09]. In chapter 4, we made use of the three most important relations in `GSAn`. Nevertheless, only the *is_a* relationship was taken into account in the semantic similarity measures. Lord et al. [Lor+03] developed an IC measure inspired by the one of Resnik [Res95] but adapted to GO. This *IC* combines, indistinctly, the *is_a* and *part_of* relationships. This assumption is delicate since combining different structures in this way may generate inconsistencies in the computation of the similarity. Later on, Wang et al. [Wan+07] presented a new measure combining both *is_a* and *part_of* relationships and applied a weighted score to provide a greater importance to a relation over the other one. Following this research could be interesting to observe different features to be used to compute the similarity between two terms.

Additionally, to better estimate the similarity between two terms, it could also be interesting to observe the differences between them. This notion of difference has been defined as the *semantic particularity* by Bettembourg et al. [BDD14]. By comparing two sets of entities, the semantic particularity corresponds to "*the value that reflects the importance of the features that belong to the first set but not the second*". The authors made use of this notion to compute the differences between two genes. For that, the sets of terms associated with each gene (and their ancestors) are compared. This type of comparison could also be carried out for two terms in an ontology and help to estimate their similarity. Thus, if two terms share a very informative ancestor and a high value of semantic particularity at the same time, these terms would be considered as less similar than if only the semantic relatedness had been computed.

To the best of our knowledge, only two semantic similarity measures considering the graph to compute the similarity between two terms have taken into account their differences [Wan+07; Son+14]. Following this research line, which considers both similarity and particularity, may improve the comparison between two terms before the clustering stage within the `GSAn` framework.

Considering generic or specific properties of ontological concepts

`GSAn` is mainly based on GO and the underlying method takes advantage of specific properties coming from the ontology structure of GO. Then, the integration of other knowledge resources that do not have the same specific properties could result in a more limited use of

them. As an illustration, to compute the incomplete annotation in GSAn, we used the *IC* proposed by Mazandu and Mulder [MM12a]. This *IC*, defined as $IC_{GO_{universal}}$ by the authors makes use of specific features of GO. As the structure of other knowledge resources may be different (as observed in section 5.3), the choice of an *IC* measure could be various. Then, it could be interesting to study the impact of using different *ICs* (or other properties that quantify the information from an ontological concept) in GSAn. As the semantic similarity, the number of *IC* measures is large and it is tricky to choose among them. In the supplementary data of [MCM17], nine *IC* measures, including the *semantic value* proposed by Wang et al. [Wan+07], have been described. Then, a future perspective could be to propose an analytical pipeline to evaluate the impact of using different *IC* measures within GSAn, when annotating a gene set according to various knowledge resources.

In particular, the GSAn method has been designed according to the taxonomy structure of GO. As observed in section 5.3, *Reactome* is structured as a partonomy, which has not been taken into consideration when this knowledge resource has been integrated within GSAn. Therefore, a future perspective would be to explore alternatives to properly include partonomies within GSAn.

Adding steps dedicated to the integration of new resources

As presented in chapter 5, mapping strategies are relevant to find inter-relations between two knowledge resources. Then, a perspective could investigate how to take advantage of these mapped terms within GSAn. For example, the computation of a *mapping score* may give additional information concerning the number of relations a term has with other terms. In DO, when a DO term is mapped to multiple GO terms, this *mapping score* could be based on the percentage of mapped GO terms and annotated genes in the set.

Moreover, for years, many efforts have focused on ontology-based integration. Therefore, tools like Hertuda [Her12], HMatch [Cas+08], SAMBO [LT06] or ServOMap [BD12] propose to align ontologies. Hertuda is a simple *string-based* matcher connecting classes and properties over a determined threshold [Her12]. HMatch uses name and context similarity to propose mappings between classes [Cas+08]. SAMBO is a system that aligns and merges biomedical ontologies by using matchers of different types (*e.g.*, *instance-based*, *string-based*) and filtering thresholds in order to suggest an ontology alignment and check its consistency [LT06]. At last, ServOMap is a large scale ontology mapping tool supporting terminologies and ontologies defined in multiple languages and computing similarity between classes thanks to information retrieval techniques [BD12; Dia14]. Therefore, these works may be relevant and should be investigated for improving the integration step in GSAn.

At last, the GO web site provides a more complex structure of GO, called GO-PLUS. This corresponds to an enriched version of GO that includes in particular other knowledge resources as ChEBI [Mat+10], Cell Ontology [Die+16] and Uberon [Mun+12]. Their exploitation and use within GSAn could also be relevant to improve the annotation information.

Including additional information provided by the genes or gene products into GSAn

Using information related to other biological networks might be useful to enrich the biological information associated with a gene set. In this context, we can refer to the interactome or "*the whole set of molecular interactions that occur within a particular cell*" [Tre12]. The three most studied interactomes are the *gene regulatory network*, the *protein-protein interaction network* and the *metabolic network* (details of interactome network and of these three types can be found in [VCB11]). By considering the inter-relations between genes, proteins or metabolites, it could be easier to extract relevant information associated with the gene set. With this idea, some enrichment analysis tools from the *Modular Enrichment Analysis* (MEA) class introduced by Huang et al. [HSL09] make use of the gene product network. For example, EnrichNet [Gla+12] uses the gene networks, provided by resources such as STRING [Szk+10], to rank the computed over-represented terms.

Considering formal concept analysis for gene set analysis

The *formal concept analysis* (*FCA*) is a mathematical theory, presented by Ganter and Wille [GW89], that links a set of objects having a set of attributes. The input of *FCA* is a matrix (or *formal context* or simply context) whose rows represent the objects, columns represent the attributes and in which relations are boolean values. A *formal concept* (or simply, a concept) in a given context is the combination of two sets: a set of objects (O) and a set of attributes (A). The particularity of this combination is that all the objects in O contains all the attributes in A and *vice-versa*. In a given context, the concepts are connected by the *is_a* relationship and the set of concepts constitute a *concept lattice*. The lattice is a graph structure where the nodes correspond to *formal concepts* and the edges represent the *is_a* relations. This lattice is rooted by a node representing the whole set of objects, has a unique leaf node that represents the whole set of attributes, and contains between these two nodes the *formal concepts* organized according to the *formal context* matrix (Figure 6.1).

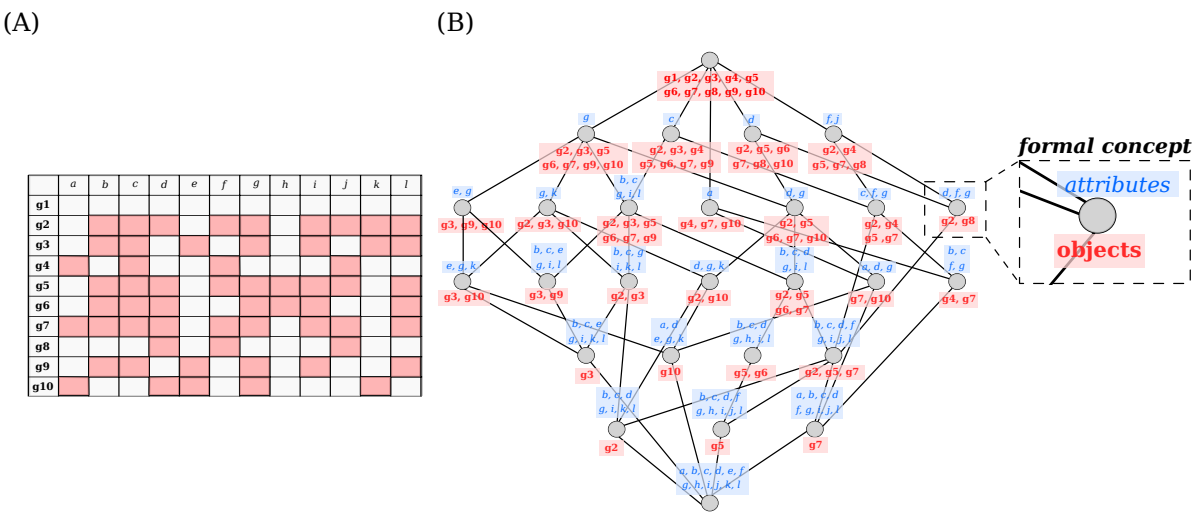


FIGURE 6.1: Illustration of formal concept analysis representation with: (A) the *formal context* describing relations between a set of objects and a set of attributes, and (B) the concept lattice represented in the graph structure.

In ontology engineering, *FCA* has often been used to design and create ontologies [Jia+03; Obi+04; Haa06; PZ07; TT12] or to merge two ontologies together [SM01; GS03; GSZ10; CBY11]. But, *FCA* is not solely focused on the creation of ontologies. *FCA* has a wide spectrum of applications (see details in [Poe+13]). In bioinformatics, most of the efforts using *FCA* are focused on gene expression or the identification of co-expressed genes [Bes+05; Bes+06; Cho+08; MVS08; Kay+09; Wer+19]. In similarity analysis, Keller et al. [KEK12] developed a semantic similarity measure based on *FCA* to compute disease similarities in a given gene set. Nevertheless, some important drawbacks must be considered [Dam16]:

- there are no relations between the elements into the set of objects,
- this analysis produces a high number of formal concepts, irrespective of whether they are non-informative or biologically irrelevant,
- *FCA* is sensible to missing or incomplete data.

As future work, I want to study the combination between *FCA* lattice structure and ontology structure in the framework of gene set annotation. For that, using the genes as objects and the ontology terms as attributes, I plan to create a lattice structure of formal concepts. Then, this structure could be used to reduce the redundancy between attributes and to associate concepts with related attributes. This re-organization of concepts would allow to be more coherent with the ontological structure or to create new concepts involving the redundant terms. Moreover, *FCA* and the lattice structure could be combined with the *MSRT* and *FCT* algorithms, presented in section 4.2.4, to improve the identification of representative terms. For that, a cluster of terms and the annotated genes for this cluster could be first used as the formal context and then used to generate the lattice structure.

6.3 Final conclusion

Nowadays, with the revolution of high-throughput technologies, such as the next generation of sequencing, the decreasing cost of sequencing makes more accessible their analysis. This accessibility allows these technologies to be routinely carried out. In this way, research fields using omics data (*e.g.*, genomic, transcriptomic, proteomics, metabolomics) have critical needs for tools dedicated to the interpretation of biological results. This thesis has contributed to the development of solutions for the functional annotation of gene sets. We focused on methods to improve the interpretation of gene sets, to propose new visualization facilities, and to integrate additional heterogeneous resources. We opened up new roads of possibilities with a novel method that combines semantic similarity measures and data mining strategies. Moreover, we investigated the visualization field, as this research domain has an important impact on biology representation and interpretation. Then, this work has demonstrated the relevance of using visual metaphors dedicated to the functional annotation analysis. Furthermore, the combination of various visual metaphors has allowed to solve limitations existing when each visual metaphor is used alone. At last, the integration of heterogeneous resources has been studied in a preliminary work but this aspect needs to be further investigated to implement a complete integration process within `GSAn`. This line of research is very interesting as the central idea is to conciliate different resources in order to extract the maximal amount of biological information.

Bibliography

[AMM44] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. "Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III". In: *Journal of experimental medicine* 79.2 (1944), pp. 137–158. DOI: 10.1097/00003086-200010001-00002.

[HC52] Alfred D Hershey and Martha Chase. "Independent functions of viral protein and nucleic acid in growth of bacteriophage". In: *The Journal of general physiology* 36.1 (1952), pp. 39–56. DOI: 10.1007/978-3-662-47150-0_3.

[FG53] Rosalind E Franklin and Raymond G Gosling. "Molecular configuration in sodium thymonucleate". In: *Nature* 171.4356 (1953), p. 740. DOI: 10.1038/171740a0.

[WC53] James D Watson and Francis HC Crick. "Molecular structure of nucleic acids". In: *Nature* 171.4356 (1953), pp. 737–738. DOI: 10.1038/171737a0.

[SR62] Robert R. Sokal and F. James Rohlf. "The comparison of dendrograms by objective methods". In: *Taxon* 11.2 (1962), pp. 33–40. ISSN: 0040-0262. DOI: 10.2307/1217208.

[Chr75] Nicos Christofides. "Graph Theory. An Algorithmic Approach". In: Computer Science and Applied Mathematics (1975). ISSN: 0884-2027. URL: https://dl.acm.org/citation.cfm?id=1098653.

[Kot80] Michael H Kottow. "A medical definition of disease". In: *Medical hypotheses* 6.2 (1980), pp. 209–213. DOI: 10.1016/0306-9877(80)90085-7.

[FM83] Edward B Fowlkes and Colin L Mallows. "A method for comparing two hierarchical clusterings". In: *Journal of the American statistical association* 78.383 (1983), pp. 553–569. DOI: 10.2307/2288117.

[Bur+85] Christian Burks, James W Fickett, Walter B Goad, Minoru Kanehisa, et al. "CABIOS REVIEW: The GenBank nucleic acid sequence database". In: *Bioinformatics* 1.4 (1985), pp. 225–233. DOI: 10.1093/bioinformatics/1.4.225.

[GW89] Bernhard Ganter and Rudolf Wille. "Conceptual Scaling". In: *Applications of Combinatorics and Graph Theory to the Biological and Social Sciences*. Ed. by Fred Roberts. New York, NY: Springer US, 1989, pp. 139–167. ISBN: 978-1-4684-6381-1. DOI: 10.1007/978-1-4684-6381-1_6.

[McC89] Alexa T. McCray. "The UMLS Semantic Network". In: *Proceedings. Symposium on Computer Applications in Medical Care* (1989), pp. 503–507. ISSN: 0195-4210. URL: http://europepmc.org/articles/PMC2245676.

[KR90] Leonard Kaufman and Peter J Rousseeuw. "Finding groups in data: An Introduction to Cluster Analysis". In: *Hoboken: Wiley Online Library* (1990). DOI: 10.2307/1269576.

[JS91] Brian Johnson and Ben Shneiderman. "Tree-Maps: A Space-filling Approach to the Visualization of Hierarchical Information Structures". In: *Proceedings of the 2Nd Conference on Visualization '91*. VIS '91. San Diego, California: IEEE Computer Society Press, 1991, pp. 284–291. ISBN: 0-8186-2245-8. URL: http://dl.acm.org/citation.cfm?id=949607.949654.

[Res95] Philip Resnik. "Using Information Content to Evaluate Semantic Similarity in a Taxonomy". In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*. IJCAI'95. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., 1995, pp. 448–453. URL: http://dl.acm.org/citation.cfm?id=1625855.1625914.

[Win+96] Edgar Wingender, Peter Dietze, Holger Karas, and Rainer Knüppel. "TRANSFAC: a database on transcription factors and their DNA binding sites". In: *Nucleic Acids Research* 24.1 (1996), pp. 238–241. DOI: 10.1093/nar/24.1.238.

[BV97] Trevor J. M. Bench-Capon and Pepijn R. S. Visser. "Ontologies in Legal Information Systems: the Need for Explicit Specifications of Domain Conceptualisations". In: *Proceedings of the 6th International Conference on Artificial Intelligence and Law*. ICAIL '97. Melbourne, Australia: ACM, 1997, pp. 132–141. ISBN: 0-89791-924-6. DOI: 10.1145/261618.261646.

[Bla+97] Judith A Blake, Joel E Richardson, Muriel T Davisson, Janan T Eppig, et al. "The Mouse Genome Database (MGD). A comprehensive public resource of genetic, phenotypic and genomic data". In: *Nucleic Acids Research* 25.1 (1997), pp. 85–91. DOI: 10.1093/nar/25.1.85.

[JC97] Jay J. Jiang and David W. Conrath. "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy". In: *Proceedings of the 10th Research on Computational Linguistics International Conference*. Taipei, Taiwan: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), 1997, pp. 19–33. URL: https://www.aclweb.org/anthology/O97-1002.pdf.

[Hel+98] Gregg A Helt, Suzanna Lewis, Ann E Loraine, and Gerald M Rubin. "BioViews: Java-based tools for genomic data visualization". In: *Genome research* 8.3 (1998), pp. 291–305. DOI: 10.1101/gr.8.3.291.

[LSW98] Ora Lassila, Ralph R. Swick, and World Wide Web Consortium. "Resource Description Framework (RDF) Model and Syntax Specification". In: (1998). URL: https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

[LC98] Claudia Leacock and Martin Chodorow. "Combining local context and WordNet similarity for word sense identification". In: *WordNet: An electronic lexical database* 49.2 (1998), pp. 265–283. DOI: 10.7551/mitpress/7287.003.0018.

[Lin98] Dekang Lin. "An Information-Theoretic Definition of Similarity". In: *Proceedings of the Fifteenth International Conference on Machine Learning*. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 296–304. ISBN: 1-55860-556-8. URL: http://dl.acm.org/citation.cfm?id=645527.657297.

[Usc+98] Mike Uschold, Martin King, Stuart Moralee, and Yannis Zorgios. "The Enterprise Ontology". In: *The Knowledge Engineering Review* 13.1 (1998), pp. 31–89. DOI: 10.1017/S0269888998001088.

[FE99] Frederico T. Fonseca and Max J. Egenhofer. "Ontology-driven Geographic Information Systems". In: *Proceedings of the 7th ACM International Symposium on Advances in Geographic Information Systems*. GIS '99. Kansas City, Missouri, USA: ACM, 1999, pp. 14–19. ISBN: 1-58113-235-2. DOI: 10.1145/320134.320137.

[JMF99] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. "Data clustering: a review". In: *ACM computing surveys (CSUR)* 31.3 (1999), pp. 264–323. DOI: 10.1145/331499.331504.

[Ash+00] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, et al. "Gene ontology: tool for the unification of biology". In: *Nature genetics* 25.1 (2000), p. 25. DOI: 10.1038/75556.

[KG00] Minoru Kanehisa and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic Acids Research* 28.1 (2000), pp. 27–30. DOI: 10.1093/nar/28.1.27.

[Lip00] Carolyn E. Lipscomb. "Medical Subject Headings (MeSH)". In: *Bulletin of the Medical Library Association* 88.3 (2000), pp. 265–266. ISSN: 0025-7338. URL: https://www.ncbi.nlm.nih.gov/pubmed/10928714.

[SGB00] Robert Stevens, Carole A Goble, and Sean Bechhofer. "Ontology-based knowledge representation for bioinformatics". In: *Briefings in bioinformatics* 1.4 (2000), pp. 398–414. DOI: 10.1093/bib/1.4.398.

[HBV01] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. "On clustering validation techniques". In: *Journal of intelligent information systems (JIIS)* 17.2-3 (2001), pp. 107–145. DOI: 10.1023/A:1012801612483.

[Lan+01] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, et al. "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822 (2001), pp. 860–921. ISSN: 1476-4687. DOI: 10.1038/35057062.

[Nis01] Darryl Nishimura. "BioCarta". In: *Biotech Software & Internet Report* 2.3 (2001), pp. 117–120. DOI: 10.1089/152791601750294344.

[Pov+01] Sue Povey, Ruth Lovering, Elspeth Bruford, Mathew Wright, et al. "The HUGO gene nomenclature committee (HGNC)". In: *Human genetics* 109.6 (2001), pp. 678–680. DOI: 10.1007/s00439-001-0615-0.

[SM01] Gerd Stumme and Alexander Maedche. "FCA-MERGE: Bottom-up Merging of Ontologies". In: *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 1*. IJCAI'01. Seattle, WA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 225–230. ISBN: 1-55860-812-5, 978-1-558-60812-2. URL: http://dl.acm.org/citation.cfm?id=1642090.1642121.

[Kar+02] Peter D Karp, Monica Riley, Milton Saier, Ian T Paulsen, et al. "The ecocyc database". In: *Nucleic Acids Research* 30.1 (2002), pp. 56–58. DOI: 10.1128/ecosalplus.esp-0006-2018.

[PS02] Viktor Pekar and Steffen Staab. "Taxonomy Learning: Factoring the Structure of a Taxonomy into a Semantic Classification Decision". In: *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*. COLING'02. Taipei, Taiwan: Association for Computational Linguistics, 2002, pp. 1–7. DOI: 10.3115/1072228.1072318.

[BVL03] Sean Bechhofer, Raphael Volz, and Phillip Lord. "Cooking the Semantic Web with the OWL API". In: *The Semantic Web - ISWC 2003*. Ed. by Dieter Fensel, Katia Sycara, and John Mylopoulos. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 659–675. ISBN: 978-3-540-39718-2. DOI: 10.1007/978-3-540-39718-2_42.

[Den+03] Glynn Dennis, Brad T Sherman, Douglas A Hosack, Jun Yang, et al. "DAVID: Database for Annotation, Visualization, and Integrated Discovery". In: *Genome biology* 4.9 (2003), R60. DOI: 10.1186/gb-2003-4-9-r60.

[Don+03] Scott W Doniger, Nathan Salomonis, Kam D Dahlquist, Karen Vranizan, et al. "MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data". In: *Genome biology* 4.1 (2003), R7. DOI: 10.1186/gb-2003-4-1-r7.

[GGW03] Prasanna Ganesan, Hector Garcia-Molina, and Jennifer Widom. "Exploiting hierarchical domain structure to compute similarity". In: *ACM Transactions on Information Systems (TOIS)* 21.1 (2003), pp. 64–93. DOI: 10.1145/635484.635487.

[GS03] Bernhard Ganter and Gerd Stumme. "Creation and Merging of Ontology Top-Levels". In: *Conceptual Structures for Knowledge Creation and Communication*. Ed. by Bernhard Ganter, Aldo de Moor, and Wilfried Lex. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 131–145. ISBN: 978-3-540-45091-7. DOI: 10.1007/978-3-540-45091-7_9.

[Gol+03] Jennifer Golbeck, Gilberto Fragoso, Frank Hartel, Jim Hendler, et al. "The National Cancer Institute's thesaurus and ontology". In: *Journal of Web Semantics First Look 1_1_4* (2003). DOI: 10.2139/ssrn.3199007.

[Huc+03] Michael Hucka, Andrew Finney, Herbert M Sauro, Hamid Bolouri, et al. "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models". In: *Bioinformatics* 19.4 (2003), pp. 524–531. DOI: 10.1093/bioinformatics/btg015.

[Jia+03] Guoqian Jiang, Katsuhiko Ogasawara, Akira Endoh, and Tsunetaro Sakurai. "Context-based ontology building support in clinical domains using formal concept analysis". In: *International journal of medical informatics* 71.1 (2003), pp. 71–81. DOI: 10.1016/s1386-5056(03)00092-3.

[KPL03] Jacob Köhler, Stephan Philippi, and Matthias Lange. "SEMEDA: ontology based semantic integration of biological databases". In: *Bioinformatics* 19.18 (2003), pp. 2420–2427. DOI: 10.1093/bioinformatics/btg340.

[Lor+03] Phillip W. Lord, Robert D. Stevens, Andy Brass, and Carole A. Goble. "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation". In: *Bioinformatics* 19.10 (2003), pp. 1275–1283. DOI: 10.1093/bioinformatics/btg153.

[Noy+03] Natalya F. Noy, Monica Crubezy, Ray W. Fergerson, Holger Knublauch, et al. "Protégé-2000: an open-source ontology-development and knowledge acquisition environment". In: *Annual Symposium proceedings. AMIA Symposium* (2003), pp. 953–953. ISSN: 1942-597X. URL: https://www.ncbi.nlm.nih.gov/pubmed/14728458.

[Rhe+03] Seung Yon Rhee, William Beavis, Tanya Z Berardini, Guanghong Chen, et al. "The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community". In: *Nucleic Acids Research* 31.1 (2003), pp. 224–228. DOI: 10.1093/nar/gkg076.

[SWS03] Barry Smith, Jennifer Williams, and Steffen Schulze-Kremer. "The ontology of the gene ontology". In: *Annual Symposium proceedings. AMIA Symposium* (2003), pp. 609–613. ISSN: 1942-597X. URL: https://www.ncbi.nlm.nih.gov/pubmed/14728245.

[Tho+03] Paul D Thomas, Michael J Campbell, Anish Kejariwal, Huaiyu Mi, et al. "PANTHER: a library of protein families and subfamilies indexed by function". In: *Genome research* 13.9 (2003), pp. 2129–2141. DOI: 10.1101/gr.772403.

[Apw+04] Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, et al. "UniProt: the universal protein knowledgebase". In: *Nucleic Acids Research* 32.suppl_1 (2004), pp. D115–D119. DOI: 10.1093/nar/gkw1099.

[BCS04] Gary D Bader, Michael P Cary, and Chris Sander. "BioPAX–biological pathway data exchange format". In: *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics* (2004). DOI: 10.1002/047001153x.g408117.

[Bec+04] Sean Bechhofer, Frank Van Harmelen, Jim Hendler, Ian Horrocks, et al. "OWL web ontology language reference". In: *W3C recommendation* 10.02 (2004). URL: https://www.w3.org/TR/owl-ref/.

[BS04] Tim Beißbarth and Terence P Speed. "GOstat: find statistically overrepresented Gene Ontologies within a group of genes". In: *Bioinformatics* 20.9 (2004), pp. 1464–1465. DOI: 10.1093/bioinformatics/bth088.

[Cam+04] Evelyn Camon, Michele Magrane, Daniel Barrell, Vivian Lee, et al. "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology". In: *Nucleic Acids Research* 32 (Jan. 2004), pp. D262–D266. ISSN: 0305-1048. DOI: 10.1093/nar/gkh021. (Visited on 05/31/2017).

[Con04a] Gene Ontology Consortium. "The Gene Ontology (GO) database and informatics resource". In: *Nucleic Acids Research* 32.suppl_1 (2004), pp. D258–D261. DOI: 10.1093/nar/gkh036.

[Con04b] International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome". In: *Nature* 431.7011 (2004), pp. 931–945. ISSN: 1476-4687. DOI: 10.1038/nature03001.

[HK04] Thomas Hernandez and Subbarao Kambhampati. "Integration of biological sources: current systems and challenges ahead". In: *ACM SIgmod Record* 33.3 (2004), pp. 51–60. DOI: 10.1145/1031570.1031583.

[JTZ04] Daxin Jiang, Chun Tang, and Aidong Zhang. "Cluster analysis for gene expression data: a survey". In: *IEEE Transactions on Knowledge and Data Engineering* 16.11 (Nov. 2004), pp. 1370–1386. ISSN: 1041-4347. DOI: 10.1109/tkde.2004.68.

[Kee04] C. Maria Keet. "Aspects of ontology integration". PhD thesis. School of Computing, Napier University, Scotland, 2004. URL: https://pdfs.semanticscholar.org/4e9d/4ccb2d011309076928a333fa21469bb24171.pdf.

[Kre+04] Lisa Kreppel, Petra Fey, Pascale Gaudet, Eric Just, et al. "dictyBase: a new Dictyostelium discoideum genome database". In: *Nucleic Acids Research* 32.suppl_1 (2004), pp. D332–D333. DOI: 10.1093/nar/gkh138.

[LHK04] Sung Geun Lee, Jung Uk Hur, and Yang Seok Kim. "A graph-theoretic modeling on GO space for biological interpretation of gene clusters". In: *Bioinformatics* 20.3 (Jan. 2004), pp. 381–388. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg420.

[Obi+04] Marek Obitko, Vaclav Snasel, Jan Smid, and V Snasel. "Ontology Design with Formal Concept Analysis." In: *CLA*. Vol. 128. 3. 2004, pp. 1377–90. URL: http://ceur-ws.org/Vol-110/paper12.pdf.

[PPM04] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. "WordNet::Similarity: Measuring the Relatedness of Concepts". In: *Demonstration Papers at HLT-NAACL 2004*. HLT-NAACL–Demonstrations '04. Boston, Massachusetts: Association for Computational Linguistics, 2004, pp. 38–41. URL: http://dl.acm.org/citation.cfm?id=1614025.1614037.

[Rah+04] Jörg Rahnenführer, Francisco S Domingues, Jochen Maydt, and Thomas Lengauer. "Calculating the statistical significance of changes in pathway activity from gene expression data". In: *Statistical applications in genetics and molecular biology* 3.1 (2004), pp. 1–29. DOI: 10.2202/1544-6115.1055.

[SVH04] Nuno Seco, Tony Veale, and Jer Hayes. "An Intrinsic Information Content Metric for Semantic Similarity in WordNet". In: *Proceedings of the 16th European Conference on Artificial Intelligence*. ECAI'04. Valencia, Spain: IOS Press, 2004, pp. 1089–1090. ISBN: 978-1-58603-452-8. URL: http://dl.acm.org/citation.cfm?id=3000001.3000272.

[SSZ04] Nora Speer, Christian Spieth, and Andreas Zell. "A memetic clustering algorithm for the functional partition of genes based on the Gene Ontology". In: *Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. Symposium on Computational Intelligence in Bioinformatics and Computational Biology. 2004, pp. 252–259. DOI: 10.1109/CIBCB.2004.1393961.

[BRA05] Jonathan Bard, Seung Y Rhee, and Michael Ashburner. "An ontology for cell types". In: *Genome biology* 6.2 (2005), R21. DOI: 10.1186/gb-2005-6-2-r21.

[Bes+05] Jérémy Besson, Céline Robardet, Jean-François Boulicaut, and Sophie Rome. "Constraint-based concept mining and its application to microarray data analysis". In: *Intelligent Data Analysis* 9.1 (2005), pp. 59–82. DOI: 10.3233/ida-2005-9105.

[Cam+05] Evelyn B Camon, Daniel G Barrell, Emily C Dimmer, Vivian Lee, et al. "An evaluation of GO annotation retrieval for BioCreAtIvE and GOA". In: *BMC bioinformatics* 6.1 (2005), S17. DOI: 10.1186/1471-2105-6-s1-s17.

[Dah05] Ralf Dahm. "Friedrich Miescher and the discovery of DNA". In: *Developmental biology* 278.2 (2005), pp. 274–288. DOI: 10.1016/j.ydbio.2004.11.028.

[Ham+05] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, et al. "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders". In: *Nucleic Acids Research* 33.suppl_1 (2005), pp. D514–D517. DOI: 10.1093/nar/gki033.

[HKK05] Julia Handl, Joshua Knowles, and Douglas B Kell. "Computational cluster validation in post-genomic data analysis". In: *Bioinformatics* 21.15 (2005), pp. 3201–3212. DOI: 10.1093/bioinformatics/bti517.

[Jos+05] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, et al. "Reactome: a knowledgebase of biological pathways". In: *Nucleic Acids Research* 33.suppl_1 (2005), pp. D428–D432. DOI: 10.1093/nar/gki072.

[LJM05] In-Yee Lee, Ho Jan-Ming, and Chen Ming-Syan. "GOMIT: a generic and adaptive annotation algorithm based on gene ontology term distributions". In: *5th IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05)*. 2005, pp. 40–48. DOI: 10.1109/BIBE.2005.33.

[Mag+05] Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. "Entrez Gene: gene-centered information at NCBI". In: *Nucleic Acids Research* 33.suppl_1 (2005), pp. D54–D58. DOI: 10.1093/nar/gkq1237.

[MR05] Oded Maimon and Lior Rokach. "Data mining and knowledge discovery handbook". In: *Database Management & Information Retrieval* (2005). DOI: 10.1007/978-0-387-09823-4.

[NW05] John C Newman and Alan M Weiner. "L2L: a simple tool for discovering the hidden significance in microarray expression data". In: *Genome biology* 6.9 (2005), R81. DOI: 10.1186/gb-2005-6-9-r81.

[RWR05] Cavan Reilly, Changchun Wang, and Mark Rutherford. "A rapid method for the comparison of cluster analyses". In: *Statistica Sinica* 15.1 (2005), pp. 19–33. ISSN: 10170405, 19968507. URL: http://www.jstor.org/stable/24307237.

[Sub+05] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, et al. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550. ISSN: 1091-6490. DOI: 10.1073/pnas.0506580102.

[Ye+05] Ping Ye, Brian D Peyser, Xuewen Pan, Jef D Boeke, et al. "Gene function prediction from congruent synthetic lethal interactions in yeast". In: *Molecular systems biology* 1.1 (2005). DOI: 10.1038/msb4100034.

[ZKS05] Bing Zhang, Stefan Kirov, and Jay Snoddy. "WebGestalt: an integrated system for exploring gene sets in various biological contexts". In: *Nucleic Acids Research* 33.suppl_2 (2005), W741–W748. DOI: 10.1093/nar/gki475.

[BCS06] Gary D Bader, Michael P Cary, and Chris Sander. "Pathguide: a pathway resource list". In: *Nucleic Acids Research* 34.suppl_1 (2006), pp. D504–D506. DOI: 10.1093/nar/gkj126.

[Bes+06] Jérémy Besson, Ruggero G. Pensa, Céline Robardet, and Jean-François Boulicaut. "Constraint-Based Mining of Fault-Tolerant Patterns from Boolean Data". In: *Knowledge Discovery in Inductive Databases*. Ed. by Francesco Bonchi and Jean-François Boulicaut. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 55–71. ISBN: 978-3-540-33293-0. DOI: 10.1007/11733492_4.

[Chi+06] Rex L Chisholm, Pascale Gaudet, Eric M Just, Karen E Pilcher, et al. "dictyBase, the model organism database for Dictyostelium discoideum". In: *Nucleic Acids Research* 34.suppl_1 (2006), pp. D423–D427. DOI: 10.1093/nar/gkj090.

[Don06] Kevin Donnelly. "SNOMED-CT: The advanced terminology and coding system for e-Health". In: *Studies in health technology and informatics* 121 (2006), p. 279. URL: http://ebooks.iospress.nl/publication/9130.

[Haa06] H. Haav. "A Semi-automatic Method to Ontology Design by using FCA". In: *Proceedings of the conference on Concept Lattices and their Applications (CLA)* (2006), pp. 13–24. URL: https://ci.nii.ac.jp/naid/20001608828/en/.

[HRM06] G Traver Hart, Arun K Ramani, and Edward M Marcotte. "How complete are current yeast and human protein-interaction networks?" In: *Genome biology* 7.11 (2006), p. 120. DOI: 10.1186/gb-2006-7-11-120.

[LT06] Patrick Lambrix and He Tan. "SAMBO—a system for aligning and merging biomedical ontologies". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 4.3 (2006), pp. 196–206. DOI: 10.2139/ssrn.3199338.

[LD06] Zhengdeng Lei and Yang Dai. "Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction". In: *BMC bioinformatics* 7.1 (2006), p. 491. DOI: 10.1186/1471-2105-7-491.

[MEG06] Glenn A Maston, Sara K Evans, and Michael R Green. "Transcriptional regulatory elements in the human genome". In: *Annu. Rev. Genomics Hum. Genet.* 7 (2006), pp. 29–59. DOI: 10.1146/annurev.genom.7.080505.115623.

[Ore+06] Eyal Oren, Knud Möller, Simon Scerri, Siegfried Handschuh, et al. "What are Semantic Annotations?" In: *Relatório técnico. DERI Galway* (2006). URL: http://www.siegfried-handschuh.net/pub/2006/whatissemannot2006.pdf.

[Riv+06] Isabelle Rivals, Léon Personnaz, Lieng Taing, and Marie-Claude Potier. "Enrichment or depletion of a GO category within a class of genes: which test?" In: *Bioinformatics* 23.4 (2006), pp. 401–407. DOI: 10.1093/bioinformatics/btl633.

[Wan+06] Weixin Wang, Hui Wang, Guozhong Dai, and Hongan Wang. "Visualization of Large Hierarchical Data by Circle Packing". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '06. New York, NY, USA: ACM, 2006, pp. 517–520. ISBN: 1-59593-372-7. DOI: 10.1145/1124772.1124851.

[Zha+06] Peisen Zhang, Jinghui Zhang, Huitao Sheng, James J. Russo, et al. "Gene functional similarity search tool (GFSST)". In: *BMC Bioinformatics* 7.1 (2006), p. 135. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-135.

[Day+07] John Day-Richter, Midori A Harris, Melissa Haendel, Gene Ontology OBO-Edit Working Group, et al. "OBO-Edit—an ontology editor for biologists". In: *Bioinformatics* 23.16 (2007), pp. 2198–2200. DOI: 10.1093/bioinformatics/btm112.

[Din+07] Li Ding, Pranam Kolari, Zhongli Ding, and Sasikanth Avancha. "Using Ontologies in the Semantic Web: A Survey". In: *Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems*. Ed. by Raj Sharman, Rajiv Kishore, and Ram Ramesh. Boston, MA: Springer US, 2007, pp. 79–113. ISBN: 978-0-387-37022-4. DOI: 10.1007/978-0-387-37022-4_4.

[Dra+07] Sorin Draghici, Purvesh Khatri, Adi Laurentiu Tarca, Kashyap Amin, et al. "A systems biology approach for pathway level analysis". In: *Genome research* 17.10 (2007), pp. 1537–1545. DOI: 10.1101/gr.6202607.

[Frö+07] Holger Fröhlich, Nora Speer, Annemarie Poustka, and Tim Beißbarth. "GOSim–an R-package for computation of information theoretic GO similarities between terms and gene products". In: *BMC bioinformatics* 8.1 (2007), p. 166. DOI: 10.1186/1471-2105-8-166.

[GK07] Martin Graham and Jessie Kennedy. "Exploring multiple trees through DAG representations". In: *IEEE Trans. on Visualization and Computer Graphics* 13.6 (2007), pp. 1294–1301. DOI: 10.1109/tvcg.2007.70556.

[Hua+07] Da W. Huang, Brad T. Sherman, Qina Tan, Joseph Kir, et al. "DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists". In: *Nucleic Acids Research* 35.suppl_2 (2007), W169–W175. DOI: 10.1093/nar/gkm415.

[Koe+07] Pierre-Yves Koenig, Guy Melançon, Charles Bohan, and Berengere Gautier. "Combining DagMaps and Sugiyama Layout for the Navigation of Hierarchical Data". In: *11th International Conference Information Visualization (IV '07)*. 2007, pp. 447–452. DOI: 10.1109/IV.2007.36.

[Mei07] Marina Meilă. "Comparing clusterings—an information based distance". In: *Journal of multivariate analysis* 98.5 (2007), pp. 873–895. DOI: 10.1016/j.jmva.2006.11.013.

[PZ07] Xin Peng and Wenyun Zhao. "An Incremental and FCA-Based Ontology Construction Method for Semantics-Based Component Retrieval". In: *7th International Conference on Quality Software*. Los Alamos, CA, USA: IEEE Computer Society, 2007, pp. 309–315. DOI: 10.1109/QSIC.2007.16.

[Pes07] Catia Pesquita. "Improving semantic similarity for proteins based on the gene ontology". MA thesis. Universidade de Lisboa, 2007. URL: https://pdfs.semanticscholar.org/5cd3/16030e26db2fb5bc1ee56869f5c2ecaab4a8.pdf.

[Rei+07] Jüri Reimand, Meelis Kull, Hedi Peterson, Jaanus Hansen, et al. "g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments". In: *Nucleic Acids Research* 35.suppl_2 (2007), W193–W200. DOI: 10.1093/nar/gkm226.

[San+07] Antonio Sanfilippo, Christian Posse, Banu Gopalan, Rick Riensche, et al. "Combining hierarchical and associative gene ontology relations with textual evidence in estimating gene and gene product similarity". In: *IEEE transactions on nanobioscience* 6.1 (2007), pp. 51–59. ISSN: 1536-1241. DOI: 10.1109/tnb.2007.891886.

[SA07] Andreas Schlicker and Mario Albrecht. "FunSimMat: a comprehensive functional similarity database". In: *Nucleic Acids Research* 36.suppl_1 (2007), pp. D434–D439. DOI: 10.1093/nar/gkm806.

[Tur+07] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, et al. "The human microbiome project". In: *Nature* 449.7164 (2007), p. 804. DOI: 10.1038/nature06244.

[Wan+07] James Z. Wang, Zhidian Du, Rapeeporn Payattakool, Philip S. Yu, et al. "A new method to measure the semantic similarity of GO terms". In: *Bioinformatics* 23.10 (2007), pp. 1274–1281. DOI: 10.1093/bioinformatics/btm087.

[Wis+07] David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, et al. "DrugBank: a knowledgebase for drugs, drug actions and drug targets". In: *Nucleic Acids Research* 36 (2007), pp. D901–D906. DOI: 10.1093/nar/gkm958.

[Bau+08] Sebastian Bauer, Steffen Grossmann, Martin Vingron, and Peter N Robinson. "Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration". In: *Bioinformatics* 24.14 (2008), pp. 1650–1651. DOI: 10.1093/bioinformatics/btn250.

[Bel+08] François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, et al. "Bio2RDF: towards a mashup to build bioinformatics knowledge systems". In: *Journal of biomedical informatics* 41.5 (2008), pp. 706–716. DOI: 10.1016/j.jbi.2008.03.004.

[But08] Atul J Butte. "Translational bioinformatics: coming of age". In: *Journal of the American Medical Informatics Association* 15.6 (2008), pp. 709–714. DOI: 10.1197/jamia.m2824.

[Cas+08] Silvana Castano, Alfio Ferrara, Davide Lorusso, and Stefano Montanelli. "The hmatch 2.0 suite for ontology matchmaking". In: *Italian Semantic Web Workshop*. Vol. 314. CEUR. 2008, pp. 1–10. URL: http://ceur-ws.org/Vol-314/36.pdf.

[Cha+08] Damien Chaussabel, Charles Quinn, Jing Shen, Pinakeen Patel, et al. "A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus". In: *Immunity* 29.1 (2008), pp. 150–164. DOI: 10.1016/j.immuni.2008.05.012.

[Cho+08] Vicky Choi, Yang Huang, Vy Lam, Dustin Potter, et al. "Using formal concept analysis for microarray data comparison". In: *Journal of bioinformatics and computational biology* 6.01 (2008), pp. 65–75. DOI: 10.1142/9781860947995_0009.

[Dav+08] Allan Peter Davis, Cynthia G Murphy, Cynthia A Saraceni-Richards, Michael C Rosenstein, et al. "Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks". In: *Nucleic Acids Research* 37 (2008), pp. D786–D792. DOI: 10.1093/nar/gkn580.

[MVS08] Susanne Motameny, Beatrix Versmold, and Rita Schmutzler. "Formal Concept Analysis for the Identification of Combinatorial Biomarkers in Breast Cancer". In: *Formal Concept Analysis*. Ed. by Raoul Medina and Sergei Obiedkov. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 229–240. ISBN: 978-3-540-78137-0. DOI: 10.1007/978-3-540-78137-0_17.

[Pic+08] Alexander R Pico, Thomas Kelder, Martijn P Van Iersel, Kristina Hanspers, et al. "WikiPathways: pathway editing for the people". In: *PLoS biology* 6.7 (2008), e184. DOI: 10.1371/journal.pbio.0060184.

[Rhe+08] Seung Yon Rhee, Valerie Wood, Kara Dolinski, and Sorin Draghici. "Use and misuse of the gene ontology annotations". In: *Nature Reviews Genetics* 9.7 (2008), p. 509. DOI: 10.1038/nrg2363.

[Rob+08] Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, et al. "The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease". In: *The American Journal of Human Genetics* 83.5 (2008), pp. 610–615. DOI: 10.1016/j.ajhg.2008.09.017.

[RSN08] D L Rubin, N H Shah, and N F Noy. "Biomedical ontologies: a functional perspective". In: *Briefings in Bioinformatics* 9.1 (2008), p. 75. DOI: 10.1093/bib/bbm059.

[XDZ08] Tao Xu, LinFang Du, and Yan Zhou. "Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data". In: *BMC bioinformatics* 9.1 (2008), p. 472. DOI: 10.1186/1471-2105-9-472.

[ZWG08a] Zili Zhou, Yanna Wang, and Junzhong Gu. "A New Model of Information Content for Semantic Similarity in WordNet". In: *2008 Second International Conference on Future Generation Communication and Networking Symposia*. Vol. 3. 2008, pp. 85–89. DOI: 10.1109/FGCNS.2008.16.

[ZWG08b] Zili Zhou, Yanna Wang, and Junzhong Gu. "New model of semantic similarity measuring in wordnet". In: *2008 3rd International Conference on Intelligent System and Knowledge Engineering*. Vol. 1. 2008, pp. 256–261. DOI: 10.1109/ISKE.2008.4730937.

[AS09] Marit Ackermann and Korbinian Strimmer. "A general modular framework for gene set enrichment analysis". In: *BMC bioinformatics* 10.1 (2009), p. 47. DOI: 10.1186/1471-2105-10-47.

[AR09] Adrian Alexa and Jörg Rahnenführer. "Gene set enrichment analysis with topGO". In: *Bioconductor Improv* 27 (2009). URL: https://bioconductor.org/packages/release/bioc/vignettes/topGO/inst/doc/topGO.pdf.

[Arn+09] Martha B Arnaud, Marcus C Chibucos, Maria C Costanzo, Jonathan Crabtree, et al. "The Aspergillus Genome Database, a curated comparative genomics resource for gene, protein and sequence information for the Aspergillus research community". In: *Nucleic Acids Research* 38.suppl_1 (2009), pp. D420–D427. DOI: 10.1093/nar/gkp751.

[Bin+09a] Gabriela Bindea, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, et al. "ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks". In: *Bioinformatics* 25.8 (2009), pp. 1091–1093. DOI: 10.1093/bioinformatics/btp101.

[Bin+09b] David Binns, Emily Dimmer, Rachael Huntley, Daniel Barrell, et al. "QuickGO: a web-based tool for Gene Ontology searching". In: *Bioinformatics* 25.22 (2009), pp. 3045–3046. DOI: 10.1093/bioinformatics/btp536.

[Car+09] Seth Carbon, Amelia Ireland, Christopher J. Mungall, ShengQiang Shu, et al. "AmiGO: online access to ontology and annotation data". In: *Bioinformatics* 25.2 (2009), pp. 288–289. DOI: 10.1093/bioinformatics/btn615.

[Du+09] Zhidian Du, Lin Li, Chin-Fu Chen, Philip S Yu, et al. "G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery". In: *Nucleic Acids Research* 37.suppl_2 (2009), W345–W349. DOI: 10.1093/nar/gkp463.

[Ede+09] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, et al. "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists". In: *BMC bioinformatics* 10.1 (2009), p. 48. DOI: 10.1186/1471-2105-10-48.

[GOS09] Nicola Guarino, Daniel Oberle, and Steffen Staab. "What is an ontology?" In: *Handbook on Ontologies*. Ed. by Steffen Staab and Rudi Studer. Heidelberg: Springer Heidelberg, 2009, pp. 1–17. ISBN: 978-3-540-92673-3. DOI: 10.1007/978-3-540-92673-3_0.

[HSL09] Da W. Huang, Brad T. Sherman, and Richard A. Lempicki. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists". In: *Nucleic Acids Research* 37.1 (2009), pp. 1–13. DOI: 10.1093/nar/gkn923.

[Kay+09] Mehdi Kaytoue, Sébastien Duplessis, Sergei O. Kuznetsov, and Amedeo Napoli. "Two FCA-Based Methods for Mining Gene Expression Data". In: *Formal Concept Analysis*. Ed. by Sébastien Ferré and Sebastian Rudolph. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 251–266. ISBN: 978-3-642-01815-2. DOI: 10.1007/978-3-642-01815-2_19.

[Pes+09] Catia Pesquita, Daniel Faria, André O. Falcão, Phillip Lord, et al. "Semantic similarity in biomedical ontologies". In: *PLOS Computational Biology* 5.7 (2009), pp. 1–12. DOI: 10.1371/journal.pcbi.1000443.

[PGC09] Catia Pesquita, Tiago Grego, and Francisco Couto. "Identifying Gene Ontology Areas for Automated Enrichment". In: *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*. Ed. by Sigeru Omatu, Miguel P. Rocha, José Bravo, Florentino Fernández, et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 934–941. ISBN: 978-3-642-02481-8. DOI: 10.1007/978-3-642-02481-8_143.

[RB09] Mark F Rogers and Asa Ben-Hur. "The use of gene ontology evidence codes in preventing classifier assessment bias". In: *Bioinformatics* 25.9 (2009), pp. 1173–1177. DOI: 10.1093/bioinformatics/btp122.

[RRN09] Troy Ruths, Derek Ruths, and Luay Nakhleh. "GS2: an efficiently computable measure of GO-based similarity of gene sets". In: *Bioinformatics* 25.9 (2009), pp. 1178–1184. DOI: 10.1093/bioinformatics/btp128.

[Sch+09] Alexandra M Schnoes, Shoshana D Brown, Igor Dodevski, and Patricia C Babbitt. "Annotation error in public databases: misannotation of molecular function in enzyme superfamilies". In: *PLoS computational biology* 5.12 (2009), e1000605. DOI: 10.1371/journal.pcbi.1000605.

[SE09] Cynthia L Smith and Janan T Eppig. "The mammalian phenotype ontology: enabling robust annotation and comparative analysis". In: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 1.3 (2009), pp. 390–399. DOI: 10.1002/wsbm.44.

[TTT09] Vassilis Tsiaras, Sofia Triantafilou, and Ioannis G. Tollis. "DAGmaps: Space Filling Visualization of Directed Acyclic Graphs". In: *Journal of Graph Algorithms and Applications* 13.3 (2009), pp. 319–347. DOI: 10.7155/jgaa.00190.

[Xu+09] Tao Xu, JianLei Gu, Yan Zhou, and LinFang Du. "Improving detection of differentially expressed gene sets by applying cluster enrichment analysis to Gene Ontology". In: *BMC bioinformatics* 10.1 (2009), p. 240. DOI: 10.1186/1471-2105-10-240.

[DSR10] Melissa J. Davis, Muhammad Shoaib B. Sehgal, and Mark A. Ragan. "Automatic, context-specific generation of Gene Ontology slims". In: *BMC Bioinformatics* 11.1 (2010), p. 498. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-498.

[GMR10] Nophar Geifman, Alon Monsonego, and Eitan Rubin. "The Neural/Immune Gene Ontology: clipping the Gene Ontology for neurological and immunological systems". In: *BMC bioinformatics* 11.1 (2010), p. 458. DOI: 10.1186/1471-2105-11-458.

[GSZ10] Li Guan-yu, Liu Shu-peng, and Zhao-Yan. "Formal concept analysis based ontology merging method". In: *3rd International Conference on Computer Science and Information Technology*. Vol. 8. 2010, pp. 279–282. DOI: 10.1109/ICCSIT.2010.5564899.

[JB10] Shobhit Jain and Gary D Bader. "An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology". In: *BMC bioinformatics* 11.1 (2010), p. 562. DOI: 10.1186/1471-2105-11-562.

[JL10] Bo Jin and Xinghua Lu. "Identifying informative subsets of the Gene Ontology with information bottleneck methods". In: *Bioinformatics (Oxford, England)* 26.19 (2010), pp. 2445–2451. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btq449.

[LG10] Jang H. Lee and Graciela H. Gonzalez. "Towards integrative gene prioritization in Alzheimer's disease". In: *Biocomputing*. 2010, pp. 4–13. DOI: 10.1142/9789814335058_0002.

[Mat+10] Paula de Matos, Adriano Dekker, Marcus Ennis, Janna Hastings, et al. "ChEBI: a chemistry ontology and database". In: *Journal of cheminformatics* 2.1 (2010), P6. DOI: 10.1186/1758-2946-2-s1-p6.

[ODo+10] Seán I O'Donoghue, Anne-Claude Gavin, Nils Gehlenborg, David S Goodsell, et al. "Visualizing biological data–now and in the future". In: *Nature methods* 7.3 (2010), S2. DOI: 10.1038/nmeth.f.301.

[SA10] Md. Hanif Seddiqui and Masaki Aono. "Metric of Intrinsic Information Content for Measuring Semantic Similarity in an Ontology". In: *Proceedings of the Seventh Asia-Pacific Conference on Conceptual Modelling - Volume 110*. APCCM '10. Brisbane, Australia: Australian Computer Society, Inc., 2010, pp. 89–96. ISBN: 978-1-920682-92-7. URL: http://dl.acm.org/citation.cfm?id=1862330.1862343.

[SB10] Sarah Song and Mik Black. "PCOT2: Principal Coordinates and Hotelling's T2 for the analysis of microarray data". In: *dim (imat)* 1.3051 (2010), p. 142. URL: http://www.bioconductor.org/packages//2.7/bioc/vignettes/pcot2/inst/doc/pcot2.pdf.

[Szk+10] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, et al. "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored". In: *Nucleic Acids Research* 39.suppl_1 (2010), pp. D561–D568. DOI: 10.1093/nar/gkq973.

[TH10] Hannah Tipney and Lawrence Hunter. "An introduction to effective use of enrichment analysis software". In: *Human genomics* 4.3 (2010), p. 202. DOI: 10.1186/1479-7364-4-3-202.

[VCH10] Lucas Vendramin, Ricardo JGB Campello, and Eduardo R Hruschka. "Relative clustering validity criteria: A comparative overview". In: *Statistical Analysis and Data Mining* 3.4 (2010), pp. 209–235. DOI: 10.1002/sam.10080.

[YS10] Genane Youness and Gilbert Saporta. "Comparing partitions of two sets of units based on the same variables". In: *Advances in data analysis and classification* 4.1 (2010), pp. 53–64. DOI: 10.1007/s11634-009-0057-4.

[Yu+10] Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, et al. "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products". In: *Bioinformatics* 26.7 (2010), pp. 976–978. DOI: 10.1093/bioinformatics/btq064.

[Ant11] Alexey V Antonov. "BioProfiling. de: analytical web portal for high-throughput cell biology". In: *Nucleic Acids Research* 39.suppl_2 (2011), W323–W327. DOI: 10.1093/nar/gkr372.

[Bak+11] Erich J Baker, Jeremy J Jay, Jason A Bubier, Michael A Langston, et al. "GeneWeaver: a web-based system for integrative functional genomics". In: *Nucleic Acids Research* 40.D1 (2011), pp. D1067–D1076. DOI: 10.1093/nar/gkr968.

[BGL11] Daniel Borcard, Francois Gillet, and Pierre Legendre. *Numerical ecology with R*. Use R! New York, NY: Springer Verlag, 2011. ISBN: 978-1-4419-7975-9. DOI: 10.1007/978-3-319-71404-2.

[BOH11] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. "D^3 data-driven documents". In: *IEEE transactions on visualization and computer graphics* 17.12 (2011), pp. 2301–2309. DOI: 10.1109/TVCG.2011.185.

[CBY11] Rung-Ching Chen, Cho-Tscan Bau, and Chun-Ju Yeh. "Merging domain ontologies based on the WordNet system and Fuzzy Formal Concept Analysis techniques". In: *Applied Soft Computing* 11.2 (2011), pp. 1908–1923. DOI: 10.1016/j.asoc.2010.06.007.

[Cul+11] Aedín C Culhane, Markus S Schröder, Razvan Sultana, Shaita C Picard, et al. "GeneSigDB: a manually curated database and resource for analysis of gene expression signatures". In: *Nucleic Acids Research* 40.D1 (2011), pp. D1060–D1066. DOI: 10.1093/nar/gkr901.

[DA11] Norberto Díaz-Díaz and Jesús S Aguilar-Ruiz. "GO-based functional dissimilarity of gene sets". In: *BMC bioinformatics* 12.1 (2011), p. 360. DOI: 10.1186/1471-2105-12-360.

[HDA11] Rachael P. Huntley, Emily C. Dimmer, and Rolf Apweiler. "Practical Applications of the Gene Ontology Resource". In: *Problem Solving Handbook in Computational Biology and Bioinformatics*. Ed. by Lenwood S. Heath and Naren Ramakrishnan. Boston, MA: Springer US, 2011, pp. 319–339. ISBN: 978-0-387-09760-2. DOI: 10.1007/978-0-387-09760-2_15.

[Jan+11] Stuart G. Jantzen, Ben JG Sutherland, David R. Minkley, and Ben F. Koop. "GO Trimming: Systematically reducing redundancy in large Gene Ontology datasets". In: *BMC Research Notes* 4.1 (2011), p. 267. ISSN: 1756-0500. DOI: 10.1186/1756-0500-4-267.

[Lam+11] Philippe Lamesch, Tanya Z Berardini, Donghui Li, David Swarbreck, et al. "The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools". In: *Nucleic Acids Research* 40.D1 (2011), pp. D1202–D1210. DOI: 10.1093/nar/gkr1090.

[Lib+11] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, et al. "Molecular signatures database (MSigDB) 3.0". In: *Bioinformatics* 27.12 (2011), pp. 1739–1740. DOI: 10.1093/bioinformatics/btr260.

[Mun+11] Christopher J. Mungall, Michael Bada, Tanya Z. Berardini, Jennifer Deegan, et al. "Cross-product extensions of the Gene Ontology". In: *Journal of biomedical informatics* 44.1 (2011), pp. 80–86. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2010.02.002.

[PŠD11] Louis du Plessis, Nives Škunca, and Christophe Dessimoz. "The what, where, how and why of gene ontology–a primer for bioinformaticians". In: *Briefings in bioinformatics* 12.6 (2011), pp. 723–735. DOI: 10.1093/bib/bbr002.

[RB11] Peter N Robinson and Sebastian Bauer. "Introduction to bio-ontologies". In: *Chapman and Hall/CRC* (2011). URL: https://www.crcpress.com/Introduction-to-Bio-Ontologies/Robinson-Bauer/p/book/9781439836651.

[SBI11] David Sánchez, Montserrat Batet, and David Isern. "Ontology-based information content computation". In: *Knowledge-Based Systems* 24.2 (2011), pp. 297–303. DOI: 10.1016/j.knosys.2010.10.001.

[Sar+11] Indra Neil Sarkar, Atul J Butte, Yves A Lussier, Peter Tarczy-Hornoch, et al. "Translational bioinformatics: linking knowledge across biological and clinical realms". In: *Journal of the American Medical Informatics Association* 18.4 (2011), pp. 354–357. DOI: 10.1136/amiajnl-2011-000245.

[Sch+11] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, et al. "Disease Ontology: a backbone for disease semantic integration". In: *Nucleic Acids Research* 40.D1 (2011), pp. D940–D946. DOI: 10.1093/nar/gkr972.

[Sch11] Hans-Jorg Schulz. "Treevis. net: A tree visualization reference". In: *IEEE Computer Graphics and Applications* 31.6 (2011), pp. 11–15. DOI: 10.1109/mcg.2011.103.

[SH11] Marek S Skrzypek and Jodi Hirschman. "Using the Saccharomyces Genome Database (SGD) for analysis of genomic information". In: *Current protocols in bioinformatics* 35.1 (2011), pp. 1–20. DOI: 10.1002/0471250953.bi0120s35.

[Sup+11] Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. "REVIGO summarizes and visualizes long lists of gene ontology terms". In: *PloS one* 6.7 (2011), e21800. DOI: 10.1371/journal.pone.0021800.

[TS11] K Thulasiraman and M N S Swamy. "Graphs: Theory and Algorithms". In: *John Wiley & Sons* (2011). URL: http://cds.cern.ch/record/1437291.

[VCB11] Marc Vidal, Michael E Cusick, and Albert-László Barabási. "Interactome networks and human disease". In: *Cell* 144.6 (2011), pp. 986–998. DOI: 10.1016/j.cell.2011.02.016.

[WL11] Xiangdong Wang and Lance Liotta. "Clinical bioinformatics: a new emerging science". In: *Journal of Clinical Bioinformatics* 1.1 (2011), p. 1. ISSN: 2043-9113. DOI: 10.1186/2043-9113-1-1.

[Zee+11] Barry R. Zeeberg, Hongfang Liu, Ari B. Kahn, Martin Ehler, et al. "RedundancyMiner: De-replication of redundant GO categories in microarray and proteomics analysis". In: *BMC bioinformatics* 12.1 (2011), p. 52. DOI: 10.1186/1471-2105-12-52.

[Abb12] Sunitha Abburu. "A survey on ontology reasoners and comparison". In: *International Journal of Computer Applications* 57.17 (2012). URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.252.5532&rep=rep1&type=pdf.

[BD12] Mouhamadou Ba and Gayo Diallo. "ServOMap and ServOMap: Lt Results for OAEI 2012". In: *Proceedings of the 7th International Conference on Ontology Matching - Volume 946*. OM'12. Boston, MA: CEUR-WS.org, 2012, pp. 197–204. URL: http://dl.acm.org/citation.cfm?id=2887596.2887615.

[Bel+12] Riccardo Bellazzi, Marco Masseroli, Shawn Murphy, Amnon Shabo, et al. "Clinical Bioinformatics: challenges and opportunities". In: *BMC Bioinformatics* 13.14 (2012), S1. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-S14-S1.

[Con+12] The ENCODE Project Consortium, Ian Dunham, Anshul Kundaje, Shelley F. Aldred, et al. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489 (2012). Article, 57 EP -. DOI: 10.1038/nature11247.

[Far+12] Daniel Faria, Andreas Schlicker, Catia Pesquita, Hugo Bastos, et al. "Mining GO annotations for improving annotation consistency". In: *PloS one* 7.7 (2012), e40519. DOI: 10.1371/journal.pone.0040519.

[Gla+12] E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, et al. "EnrichNet: network-based gene set enrichment analysis". In: *Bioinformatics* 28.18 (2012), p. i451. DOI: 10.1093/bioinformatics/bts389.

[Guz+12] Pietro H. Guzzi, Marco Mina, Concettina Guerra, and Mario Cannataro. "Semantic similarity analysis of protein data: assessment with biological features and issues". In: *Briefings in Bioinformatics* 13.5 (2012), pp. 569–585. ISSN: 1477-4054. DOI: 10.1093/bib/bbr066.

[Her12] Sven Hertling. "Hertuda Results for OEAI 2012". In: *Proceedings of the 7th International Conference on Ontology Matching - Volume 946*. OM'12. Boston, MA: CEUR-WS.org, 2012, pp. 141–144. URL: http://dl.acm.org/citation.cfm?id=2887596.2887607.

[Jia+12] Xiaoli Jiao, Brad T Sherman, Da Wei Huang, Robert Stephens, et al. "DAVID-WS: a stateful web service to facilitate gene/protein list analysis". In: *Bioinformatics* 28.13 (2012), pp. 1805–1806. DOI: 10.1093/bioinformatics/bts251.

[KH12] Johannes Kehrer and Helwig Hauser. "Visualization and visual analysis of multifaceted scientific data: A survey". In: *IEEE transactions on visualization and computer graphics* 19.3 (2012), pp. 495–513. DOI: 10.1109/tvcg.2012.110.

[KEK12] Benjamin J. Keller, Felix Eichinger, and Matthias Kretzler. "Formal concept analysis of disease similarity". In: *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* 2012 (2012). 22779049[pmid], pp. 42–51. ISSN: 2153-4063. URL: https://www.ncbi.nlm.nih.gov/pubmed/22779049.

[KSB12] Purvesh Khatri, Marina Sirota, and Atul J Butte. "Ten years of pathway analysis: current approaches and outstanding challenges". In: *PLoS computational biology* 8.2 (2012), e1002375. DOI: 10.1371/journal.pcbi.1002375.

[MBV12] Carmen Martinez-Cruz, Ignacio J Blanco, and M Amparo Vila. "Ontologies versus relational databases: are they so different? A comparison". In: *Artificial Intelligence Review* 38.4 (2012), pp. 271–290. DOI: 10.1007/s10462-011-9251-9.

[MM12a] Gaston K. Mazandu and Nicola J. Mulder. "A topology-based metric for measuring term similarity in the Gene Ontology". In: *Advances in Bioinformatics* 2012 (2012). ISSN: 1687-8027. DOI: 10.1155/2012/975783. (Visited on 04/13/2018).

[MM12b] Gaston K Mazandu and Nicola J Mulder. "Function prediction and analysis of Mycobacterium tuberculosis hypothetical proteins". In: *International journal of molecular sciences* 13.6 (2012), pp. 7283–7302. DOI: 10.3390/ijms13067283.

[MZ12] Isabella Morlini and Sergio Zani. "Dissimilarity and similarity measures for comparing dendrograms and their applications". In: *Advances in Data Analysis and Classification* 6.2 (2012), pp. 85–105. DOI: 10.1007/s11634-012-0106-2.

[Mun+12] Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, et al. "Uberon, an integrative multi-species anatomy ontology". In: *Genome biology* 13.1 (2012), R5. DOI: 10.1186/gb-2012-13-1-r5.

[Sch+12] Daniel J Schaid, Jason P Sinnwell, Gregory D Jenkins, Shannon K McDonnell, et al. "Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies". In: *Genetic epidemiology* 36.1 (2012), pp. 3–16. DOI: 10.1002/gepi.20632.

[TT12] Chien D. C. Ta and Tuoi P. Thi. "Improving the formal concept analysis algorithm to construct domain ontology". In: *2012 Fourth International Conference on Knowledge and Systems Engineering*. 2012, pp. 74–78. DOI: 10.1109/KSE.2012.42.

[TNP12] Daniel Tabas-Madrid, Ruben Nogales-Cadenas, and Alberto Pascual-Montano. "GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics". In: *Nucleic Acids Research* 40.W1 (2012), W478–W483. DOI: 10.1093/nar/gks402.

[Tre12] Ronald J Trent. "Chapter 4 - Omics". In: *Molecular Medicine (Fourth Edition)*. Ed. by Ronald J Trent. Fourth Edition. Boston/Waltham: Academic Press, 2012, pp. 117–152. ISBN: 978-0-12-381451-7. DOI: 10.1016/B978-0-12-381451-7.00004-9.

[Yu+12] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. "clusterProfiler: an R package for comparing biological themes among gene clusters". In: *Omics: a journal of integrative biology* 16.5 (2012), pp. 284–287. DOI: 10.1089/omi.2011.0118.

[Bin+13] Gabriela Bindea, Bernhard Mlecnik, Marie Tosolini, Amos Kirilovsky, et al. "Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer". In: *Immunity* 39.4 (2013), pp. 782–795. DOI: 10.1016/j.immuni.2013.10.003.

[Che+13] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, et al. "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool". In: *BMC bioinformatics* 14.1 (2013), p. 128. DOI: 10.1186/1471-2105-14-128.

[FHC13] João D. Ferreira, Janna Hastings, and Francisco M. Couto. "Exploiting disjointness axioms to improve semantic similarity measures". In: *Bioinformatics* 29.21 (2013), pp. 2781–2787. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt491.

[FNS13] Bo Fu, Natalya F. Noy, and Margaret-Anne Storey. "Indented Tree or Graph? A Usability Study of Ontology Visualization Techniques in the Context of Class Mapping Evaluation". In: *The Semantic Web – ISWC 2013*. Ed. by Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 117–134. ISBN: 978-3-642-41335-3. DOI: 10.1007/978-3-642-41335-3_8.

[Har+13] Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. "The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies". In: *Bioinformatics* 30.5 (2013), pp. 740–742. DOI: 10.1093/bioinformatics/btt581.

[LTP13] Weil Lai, Lu Tian, and Peter Park. "sigPathway: Pathway Analysis with Microarray Data". In: (2013). DOI: 10.18129/B9.bioc.sigPathway.

[Li+13] Shuzhao Li, Nadine Rouphael, Sai Duraisingham, Sandra Romero-Steiner, et al. "Molecular signatures of antibody responses derived from a systems biology study of five human vaccines". In: *Nature Immunology* 15 (2013), pp. 195–204. DOI: 10.1038/ni.2789.

[MMB13] Prashanti Manda, Fiona McCarthy, and Susan M. Bridges. "Interestingness measures and strategies for mining multi-ontology multi-level association rules from gene ontology annotations for the discovery of new GO relationships". In: *Journal of Biomedical Informatics* 46.5 (2013), pp. 849–856. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2013.06.012. URL: http://www.sciencedirect.com/science/article/pii/S1532046413000877.

[MM13] Gaston K. Mazandu and Nicola J. Mulder. "Information content-based Gene Ontology semantic similarity approaches: toward a unified framework theory". eng. In: *BioMed Research International* 2013 (2013), p. 292063. ISSN: 2314-6141. DOI: 10.1155/2013/292063.

[Nau+13] Stefan Naulaerts, Pieter Meysman, Wout Bittremieux, Trung Nghia Vu, et al. "A primer to frequent itemset mining for bioinformatics". In: *Briefings in bioinformatics* 16.2 (2013), pp. 216–231. DOI: 10.1093/bib/bbt074.

[Nev+13] Kornelia Neveling, Ilse Feenstra, Christian Gilissen, Lies H Hoefsloot, et al. "A Post-Hoc Comparison of the Utility of S anger Sequencing and Exome Sequencing for the Diagnosis of Heterogeneous Diseases". In: *Human mutation* 34.12 (2013), pp. 1721–1726. DOI: 10.1002/humu.22450.

[Paf+13] Evangelos Pafilis, Sune P Frankild, Lucia Fanini, Sarah Faulwetter, et al. "The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text". In: *PLoS One* 8.6 (2013), pp. 1–6. DOI: 10.1371/journal.pone.0065390.

[Poe+13] Jonas Poelmans, Dmitry I Ignatov, Sergei O Kuznetsov, and Guido Dedene. "Formal concept analysis in knowledge processing: A survey on applications". In: *Expert systems with applications* 40.16 (2013), pp. 6538–6560. DOI: 10.1016/j.eswa.2013.05.009.

[Pri+13] CR Primmer, S Papakostas, EH Leder, MJ Davis, et al. "Annotated genes and nonannotated genomes: cross-species use of Gene Ontology in ecology and evolution research". In: *Molecular ecology* 22.12 (2013), pp. 3216–3241. DOI: 10.1111/mec.12309.

[SE13] P. Shvaiko and J. Euzenat. "Ontology Matching: State of the Art and Future Challenges". In: *IEEE Transactions on Knowledge and Data Engineering* 25.1 (2013), pp. 158–176. DOI: 10.1109/TKDE.2011.253.

[Xu+13] Yungang Xu, Maozu Guo, Wenli Shi, Xiaoyan Liu, et al. "A novel insight into Gene Ontology semantic similarity". In: *Genomics* 101.6 (2013), pp. 368–375. ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2013.04.010.

[ALP14] Abul K Abbas, Andrew HH Lichtman, and Shiv Pillai. "Cellular and molecular immunology". In: *Elsevier Health Sciences* (2014). URL: https://www.elsevier.com/books/cellular-and-molecular-immunology/abbas/978-0-323-47978-3.

[Bec+14] Fabian Beck, Michael Burch, Stephan Diehl, and Daniel Weiskopf. "The State of the Art in Visualizing Dynamic Graphs". In: *EuroVis - STARs*. Ed. by R. Borgo, R. Maciejewski, and I. Viola. The Eurographics Association, 2014. ISBN: 978-3-03868-028-4. DOI: 10.2312/eurovisstar.20141174.

[BSS14] Fernando Benites, Svenja Simon, and Elena Sapozhnikova. "Mining rare associations between biological ontologies". In: *PloS one* 9.1 (2014), e84475. DOI: 10.1371/journal.pone.0084475.

[BDD14] Charles Bettembourg, Christian Diot, and Olivier Dameron. "Semantic particularity measure for functional characterization of gene sets using gene ontology". In: *PloS one* 9.1 (2014), e86525. DOI: 10.1371/journal.pone.0086525.

[Bra+14] A Bravo, M Cases, N Queralt-Rosinach, F Sanz, et al. "A knowledge-driven approach to extract disease-related biomarkers from the literature". In: *BioMed research international* 2014 (2014). DOI: 10.1155/2014/253128.

[CB14] Damien Chaussabel and Nicole Baldwin. "Democratizing systems immunology with modular transcriptional repertoire analyses". In: *Nature Reviews Immunology* 14.4 (2014), p. 271. DOI: 10.1038/nri3642.

[Dia14] Gayo Diallo. "An effective method of large scale ontology matching". In: *Journal of biomedical semantics* 5.1 (2014), p. 44. DOI: 10.1186/2041-1480-5-44.

[GMC14] Pietro Hiram Guzzi, Marianna Milano, and Mario Cannataro. "Mining Association Rules from Gene Ontology and Protein Networks: Promises and Challenges." In: *Procedia Computer Science* 29 (2014). 2014 International Conference on Computational Science, pp. 1970–1980. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2014.05.181.

[HAZ14] Joke M.M. den Haan, Ramon Arens, and Menno C. van Zelm. "The activation of the adaptive immune system: Cross-talk between antigen presenting cells, T cells and B cells". In: *Immunology Letters* 162.2, Part B (2014), pp. 103–112. ISSN: 0165-2478. DOI: 10.1016/j.imlet.2014.10.011.

[Hun+14] Rachael P Huntley, Tony Sawford, Maria J Martin, and Claire O'Donovan. "Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt". In: *GigaScience* 3.1 (2014), p. 4. DOI: 10.1186/2047-217x-3-4.

[KL14] Taku Kambayashi and Terri M. Laufer. "Atypical MHC class II-expressing antigen presenting cells: can anything replace a dendritic cell?" In: *Nature Reviews Immunology* 14 (2014). Review Article, p. 719. DOI: 10.1038/nri3754.

[LW14] Dao Lam and Donald C. Wunsch. "Chapter 20 - Clustering". In: *Academic Press Library in Signal Processing: Volume 1*. Ed. by Paulo S.R. Diniz, Johan A.K. Suykens, Rama Chellappa, and Sergios Theodoridis. Vol. 1. Academic Press Library in Signal Processing. Elsevier, 2014, pp. 1115–1149. DOI: 10.1016/B978-0-12-396502-8.00020-6.

[Lib14] Arthur Liberzon. "A Description of the Molecular Signatures Database (MSigDB) Web Site". In: *Stem Cell Transcriptional Networks: Methods and Protocols*. Ed. by Benjamin L. Kidder. New York, NY: Springer New York, 2014, pp. 153–160. ISBN: 978-1-4939-0512-6. DOI: 10.1007/978-1-4939-0512-6_9.

[Man+14] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, et al. "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 55–60. DOI: 10.3115/v1/P14-5010.

[MM14] Gaston K. Mazandu and Nicola J. Mulder. "Information content-based Gene Ontology functional similarity measures: which one to use for a given biological data type?" en. In: *PLOS ONE* 9.12 (2014), e113859. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0113859.

[NSG14] Dokyun Na, Hyungbin Son, and Jörg Gsponer. "Categorizer: a tool to categorize genes into user-defined biological groups based on semantic similarity". In: *BMC Genomics* 15 (2014), p. 1091. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-1091.

[Nun+14] Teresa Nunez Villavicencio-Diaz, Arielis Rodríguez-Ulloa, Osmany Guirola-Cruz, and Yasset Perez-Riverol. "Bioinformatics tools for the functional interpretation of quantitative proteomics results". In: *Current topics in medicinal chemistry* 14.3 (2014), pp. 435–449. DOI: 10.2174/1568026613666131204105110.

[PD14] Jennifer L Plank and Ann Dean. "Enhancer function: mechanistic and genome-wide insights come together". In: *Molecular cell* 55.1 (2014), pp. 5–14. DOI: 10.1016/j.molcel.2014.06.015.

[Sla14] Ted Slater. "Recent advances in modeling languages for pathway maps and computable biological networks". In: *Drug discovery today* 19.2 (2014), pp. 193–198. DOI: 10.1016/j.drudis.2013.12.011.

[Son+14] Xuebo Song, Lin Li, Pradip K. Srimani, and James Z. Wang. "Measure the semantic similarity of GO terms using Aggregate Information Content". In: *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 11.3 (2014), pp. 468–476. ISSN: 1545-5963. DOI: 10.1109/TCBB.2013.176.

[Szk+14] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, et al. "STRING v10: protein–protein interaction networks, integrated over the tree of life". In: *Nucleic Acids Research* 43.D1 (2014), pp. D447–D452. DOI: 10.1093/nar/gku1003.

[TJ14] M. Tennekes and E. de Jonge. "Tree colors: Color schemes for tree-structured data". In: *IEEE trans. on visualization and computer graphics* 20.12 (2014), pp. 2072–2081. DOI: 10.1109/tvcg.2014.2346277.

[Vas+14] Drashtti Vasant, Laetitia Chanas, James Malone, Marc Hanauer, et al. "ORDO: An ontology connecting rare disease, epidemiology and genetic data". In: *Proceedings of ISMB*. Vol. 30. 2014. URL: https://www.researchgate.net/publication/281824026_ORDO_An_Ontology_Connecting_Rare_Disease_Epidemiology_and_Genetic_Data.

[Aga+15] Giuseppe Agapito, Mario Cannataro, Pietro Hiram Guzzi, and Marianna Milano. "Using GO-WAR for mining cross-ontology weighted association rules". In: *Computer Methods and Programs in Biomedicine* 120.2 (2015), pp. 113–122. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2015.03.007.

[Bak+15] Erich Baker, Jason A Bubier, Timothy Reynolds, Michael A Langston, et al. "GeneWeaver: data driven alignment of cross-species genomics in biology and disease". In: *Nucleic Acids Research* 44.D1 (2015), pp. D555–D559. DOI: 10.1093/nar/gkv1329.

[Bel+15] Frida Belinky, Noam Nativ, Gil Stelzer, Shahar Zimmerman, et al. "PathCards: multi-source consolidation of human biological pathways". In: *Database* (2015). DOI: 10.1093/database/bav006.

[BLG15] Thomas Bleazard, Janine A Lamb, and Sam Griffiths-Jones. "Bias in microRNA functional enrichment analysis". In: *Bioinformatics* 31.10 (2015), p. 1592. DOI: 10.1093/bioinformatics/btv023.

[Bra+15] Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, et al. "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research". In: *BMC bioinformatics* 16.1 (2015), p. 55. DOI: 10.1101/007443.

[Che15] J Michael Cherry. "The Saccharomyces genome database: a tool for discovery". In: *Cold Spring Harbor Protocols* 2015.12 (2015), pdb–top083840. DOI: 10.1101/pdb.top083840.

[Con+15] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), p. 68. DOI: 10.1038/nature15393.

[GB15] Romain Giot and Romain Bourqui. "Fast Graph Drawing Algorithm Revealing Networks Cores". In: *2015 19th International Conference on Information Visualisation*. 2015, pp. 259–264. DOI: 10.1109/iV.2015.54.

[Har+15] Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. "Semantic similarity from natural language and ontology analysis". In: *Morgan & Claypool Publishers* 8.1 (2015). DOI: 10.2200/s00639ed1v01y201504hlt027.

[HL15] Oren Harman and Ehud Lamm. "History of Classical Genetics". In: *eLS*. American Cancer Society, 2015, pp. 1–6. DOI: 10.1002/9780470015902.a0003094.pub2.

[Hen15] Christian Hennig. "What are the true clusters?" In: *Pattern Recognition Letters* 64 (2015). Philosophical Aspects of Pattern Recognition, pp. 53–62. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2015.04.009.

[HSG15] Robert Hoehndorf, Paul N. Schofield, and Georgios V. Gkoutos. "The role of ontologies in biological and biomedical research: a functional perspective". In: *Briefings in Bioinformatics* 16.6 (2015), p. 1069. DOI: 10.1093/bib/bbv011.

[Kha+15] Afroza Khanam Irin, Alpha Tom Kodamullil, Michaela Gündel, and Martin Hofmann-Apitius. "Computational modelling approaches on epigenetic factors in neurodegenerative and autoimmune diseases and their mechanistic analysis". In: *Journal of immunology research* (2015). DOI: 10.1155/2015/737168.

[KK15] Kostiantyn Kucher and Andreas Kerren. "Text visualization techniques: Taxonomy, visual survey, and community insights". In: *2015 IEEE Pacific Visualization Symposium (PacificVis)*. 2015, pp. 117–121. DOI: 10.1109/PACIFICVIS.2015.7156366.

[Maz+15] Gaston K Mazandu, Emile R Chimusa, Mamana Mbiyavanga, and Nicola J Mulder. "A-DaGO-Fun: an adaptable Gene Ontology semantic similarity-based functional analysis tool". In: *Bioinformatics* 32.3 (2015), pp. 477–479. DOI: 10.1093/bioinformatics/btv590.

[NKH15] Mufassra Naz, Alpha Tom Kodamullil, and Martin Hofmann-Apitius. "Reasoning over genetic variance information in cause-and-effect models of neurodegenerative diseases". In: *Briefings in bioinformatics* 17.3 (2015), pp. 505–516. DOI: 10.1093/bib/bbv063.

[OLe+15] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". In: *Nucleic Acids Research* 44.D1 (2015), pp. D733–D745. DOI: 10.1093/nar/gkv1189.

[ORG15] Lorena Otero-Cerdeira, Francisco J Rodríguez-Martínez, and Alma Gómez-Rodríguez. "Ontology matching: A literature review". In: *Expert Systems with Applications* 42.2 (2015), pp. 949–971. DOI: 10.1016/j.eswa.2014.08.032.

[Pav+15] Georgios A Pavlopoulos, Dimitris Malliarakis, Nikolas Papanikolaou, Theodosis Theodosiou, et al. "Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future". In: *Gigascience* 4.1 (2015), p. 38. DOI: 10.1186/s13742-015-0077-2.

[PA15] Alexandre Perrot and David Auber. "FATuM - Fast Animated Transitions Using Multi-buffers". In: *2015 19th International Conference on Information Visualisation*. 2015, pp. 75–82. DOI: 10.1109/iV.2015.24.

[Pil+15] Eleftherios Pilalis, Theodoros Koutsandreas, Ioannis Valavanis, Emmanouil Athanasiadis, et al. "KENeV: A web-application for the automated reconstruction and visualization of the enriched metabolic and signaling super-pathways deriving from genomic experiments". In: *Computational and structural biotechnology journal* 13 (2015), pp. 248–255. DOI: 10.1016/j.csbj.2015.03.009.

[Piñ+15] Janet Piñero, Núria Queralt-Rosinach, Alex Bravo, Jordi Deu-Pons, et al. "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes". In: *Database* 2015 (2015). DOI: 10.1093/database/bav028.

[Ple+15] Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafou, Janos X Binder, et al. "DISEASES: Text mining and data integration of disease–gene associations". In: *Methods* 74 (2015), pp. 83–89. DOI: 10.1101/008425.

[Que+15a] Manuel Quesada-Martínez, Jesualdo Tomás Fernández-Breis, Robert Stevens, and Nathalie Aussenac-Gilles. "OntoEnrich: A Platform for the Lexical Analysis of Ontologies". In: *Knowledge Engineering and Knowledge Management*. Ed. by Patrick Lambrix, Eero Hyvönen, Eva Blomqvist, Valentina Presutti, et al. Cham: Springer International Publishing, 2015, pp. 172–176. ISBN: 978-3-319-17966-7. DOI: 10.1007/978-3-319-17966-7_25.

[Que+15b] Manuel Quesada-Martínez, Jesualdo Tomás Fernández-Breis, Robert Stevens, and Eleni Mikroyannidi. "Prioritising lexical patterns to increase axiomatisation in biomedical ontologies". In: *Methods of information in medicine* 54.1 (2015), pp. 56–64. DOI: 10.3414/me13-02-0026.

[Shi+15] Youhyun Shin, Yeonchan Ahn, Heesik Jeon, and Sang-goo Lee. "Consensus Similarity Measure for Short Text Clustering". In: *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*. 2015, pp. 264–268. DOI: 10.1109/DEXA.2015.65.

[The15] The Gene Ontology Consortium. "Gene Ontology Consortium: going forward". In: *Nucleic Acids Research* 43.Database issue (2015), pp. D1049–D1056. ISSN: 0305-1048. DOI: 10.1093/nar/gku1179.

[Thé+15] Patricia Thébault, Romain Bourqui, William Benchimol, Christine Gaspin, et al. "Advantages of mixing bioinformatics and visualization approaches for analyzing sRNA-mediated regulatory bacterial networks". In: *Briefings in Bioinformatics* 16.5 (2015), pp. 795–805. ISSN: 1477-4054. DOI: 10.1093/bib/bbu045.

[VBW15] Corinna Vehlow, Fabian Beck, and Daniel Weiskopf. "The State of the Art in Visualizing Group Structures in Graphs". In: *Eurographics Conference on Visualization (EuroVis) - STARs*. Ed. by R. Borgo, F. Ganovelli, and I. Viola. The Eurographics Association, 2015. DOI: 10.2312/eurovisstar.20151110.

[ZL15] Haisen Zhao and Lin Lu. "Variational circular treemaps for interactive visualization of hierarchical data". In: *2015 IEEE Pacific Visualization Symposium (PacificVis)*. 2015, pp. 81–85. DOI: 10.1109/PACIFICVIS.2015.7156360.

[Alh+16] Monther Alhamdoosh, Milica Ng, Nicholas J Wilson, Julie M Sheridan, et al. "Combining multiple tools outperforms individual methods in gene set enrichment analyses". In: *Bioinformatics* 33.3 (2016), pp. 414–424. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw623.

[BPG16] Sara Ballouz, Paul Pavlidis, and Jesse Gillis. "Using predictive specificity to determine when gene set analysis is biologically meaningful". In: *Nucleic Acids Research* 45.4 (2016), e20–e20. DOI: 10.1101/080127.

[Ben+16] Shani Ben-Ari Fuchs, Iris Lieder, Gil Stelzer, Yaron Mazor, et al. "GeneAnalytics: an integrative gene set analysis tool for next generation sequencing, RNAseq and microarray data". In: *Omics: a journal of integrative biology* 20.3 (2016), pp. 139–151. DOI: 10.1089/omi.2015.0168.

[Cou+16] Mélanie Courtot, Alex L Mitchell, Maxim Scheremetjew, Janet Piñero González, et al. "Slim-o-matic: a Semi-Automated Way to Generate Gene Ontology Slims." In: *SWAT4LS* 2016 (2016). URL: http://www.swat4ls.org/wp-content/uploads/2016/10/paper-23.pdf.

[Dam16] Olivier Dameron. "Ontology-based methods for analyzing life science data". PhD thesis. Université de Rennes, 2016. URL: https://hal.inria.fr/tel-01403371.

[Die+16] Alexander D Diehl, Terrence F Meehan, Yvonne M Bradford, Matthew H Brush, et al. "The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability". In: *Journal of biomedical semantics* 7.1 (2016), p. 44. DOI: 10.1186/s13326-016-0088-7.

[Man+16] Claudia Manzoni, Demis A Kia, Jana Vandrovcova, John Hardy, et al. "Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences". In: *Briefings in bioinformatics* 19.2 (2016), pp. 286–302. DOI: 10.1093/bib/bbw114.

[Mar16] Jorge Martinez-Gil. "CoTO: A novel approach for fuzzy aggregation of semantic similarity measures". In: *Cognitive Systems Research* 40 (2016), pp. 8–17. ISSN: 1389-0417. DOI: 10.1016/j.cogsys.2016.01.001.

[MCG16] Steven J. Marygold, Madeline A. Crosby, and Joshua L. Goodman. "Using FlyBase, a Database of Drosophila Genes and Genomes". In: *Drosophila: Methods and Protocols*. Ed. by Christian Dahmann. New York, NY: Springer New York, 2016, pp. 1–31. ISBN: 978-1-4939-6371-3. DOI: 10.1007/978-1-4939-6371-3_1.

[Mıs+16] Göksel Mısırlı, Jennifer Hallinan, Matthew Pocock, Phillip Lord, et al. "Data integration and mining for synthetic biology design". In: *ACS synthetic biology* 5.10 (2016), pp. 1086–1097. DOI: 10.1021/acssynbio.5b00295.

[PK16] Sang Tae Park and Jayoung Kim. "Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing". In: *International Neurourology Journal* 20 (Suppl 2 2016), S76–83. ISSN: 2093-4777. DOI: 10.5213/inj.1632742.371.

[Rei+16] Jüri Reimand, Tambet Arak, Priit Adler, Liis Kolberg, et al. "g:Profiler–a web server for functional interpretation of gene lists (2016 update)". In: *Nucleic Acids Research* 44.W1 (2016), W83–W89. DOI: 10.1093/nar/gkw199.

[SVW16] Mirko Signorelli, Veronica Vinciotti, and Ernst C Wit. "NEAT: an efficient network enrichment analysis test". In: *BMC bioinformatics* 17.1 (2016), p. 352. DOI: 10.1186/s12859-016-1203-6.

[The16] The UniProt Consortium. "UniProt: the universal protein knowledgebase". In: *Nucleic Acids Research* 45.D1 (2016), pp. D158–D169. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1099.

[VLZ16] Francis J. Vasko, Yun Lu, and Kenneth Zyma. "What is the best greedy-like heuristic for the weighted set covering problem?" In: *Operations Research Letters* 44.3 (2016), pp. 366–369. ISSN: 0167-6377. DOI: 10.1016/j.orl.2016.03.007.

[Emo+17] Mohammad Asif Emran Khan Emon, Reagon Karki, Erfan Younesi, Martin Hofmann-Apitius, et al. "Using drugs as molecular probes: A computational chemical biology approach in neurodegenerative diseases". In: *Journal of Alzheimer's Disease* 56.2 (2017), pp. 677–686. DOI: 10.3233/jad-160222.

[FNS17] Bo Fu, Noy Noy, and Margaret-Anne Storey. "Eye tracking the user experience–An evaluation of ontology visualization techniques". In: *Semantic Web* 8.1 (2017), pp. 23–41. DOI: 10.3233/sw-140163.

[GD17] Pascale Gaudet and Christophe Dessimoz. "Gene Ontology: Pitfalls, Biases, and Remedies". In: *The Gene Ontology Handbook*. Ed. by Christophe Dessimoz and Nives Škunca. New York, NY: Springer New York, 2017, pp. 189–205. ISBN: 978-1-4939-3743-1. DOI: 10.1007/978-1-4939-3743-1_14.

[Gau+17] Pascale Gaudet, Nives Škunca, James C. Hu, and Christophe Dessimoz. "Primer on the Gene Ontology". In: *The Gene Ontology Handbook*. Ed. by Christophe Dessimoz and Nives Škunca. New York, NY: Springer New York, 2017, pp. 25–37. ISBN: 978-1-4939-3743-1. DOI: 10.1007/978-1-4939-3743-1_3.

[Has17] Janna Hastings. "Primer on Ontologies". In: *The Gene Ontology Handbook*. Ed. by Christophe Dessimoz and Nives Škunca. New York, NY: Springer New York, 2017, pp. 3–13. ISBN: 978-1-4939-3743-1. DOI: 10.1007/978-1-4939-3743-1_1.

[Iya+17] Anandhi Iyappan, Erfan Younesi, Alberto Redolfi, Henri Vrooman, et al. "Neuroimaging feature terminology: A controlled terminology for the annotation of brain imaging features". In: *Journal of Alzheimer's Disease* 59.4 (2017), pp. 1153–1169. DOI: 10.3233/jad-161148.

[Ker+17] Andreas Kerren, Kostiantyn Kucher, Yuan-Fang Li, and Falk Schreiber. "BioVis Explorer: A visual guide for biological data visualization techniques". In: *PloS one* 12.11 (2017), e0187341. DOI: 10.1371/journal.pone.0187341.

[MCM17] Gaston K. Mazandu, Emile R. Chimusa, and Nicola J. Mulder. "Gene Ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery". In: *Briefings in Bioinformatics* 18.5 (2017), pp. 886–901. DOI: 10.1093/bib/bbw067.

[MC17] Monica Munoz-Torres and Seth Carbon. "Get GO! Retrieving GO Data Using AmiGO, QuickGO, API, Files, and Tools". In: *The Gene Ontology Handbook*. Ed. by Christophe Dessimoz and Nives Škunca. New York, NY: Springer New York, 2017, pp. 149–160. ISBN: 978-1-4939-3743-1. DOI: 10.1007/978-1-4939-3743-1_11.

[SŠ17] Fran Supek and Nives Škunca. "Visualizing GO Annotations". In: *The Gene Ontology Handbook*. Ed. by Christophe Dessimoz and Nives Škunca. New York, NY: Springer New York, 2017, pp. 207–220. ISBN: 978-1-4939-3743-1. DOI: 10.1007/978-1-4939-3743-1_15.

[Tho17] Paul D. Thomas. "The Gene Ontology and the Meaning of Biological Function". In: *The Gene Ontology Handbook*. Ed. by Christophe Dessimoz and Nives Škunca. New York, NY: Springer New York, 2017, pp. 15–24. ISBN: 978-1-4939-3743-1. DOI: 10.1007/978-1-4939-3743-1_2.

[Yan+17] Jingwen Yan, Shannon L Risacher, Li Shen, and Andrew J Saykin. "Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data". In: *Briefings in bioinformatics* 19.6 (2017), pp. 1370–1381. DOI: 10.1093/bib/bbx066.

[ZYL17] Tao Zhou, Jun Yao, and Zhanjiang Liu. "Gene ontology, enrichment analysis, and pathway analysis". In: *Bioinformatics in aquaculture: Principles and methods* (2017), pp. 150–168. DOI: 10.1002/9781118782392.ch10.

[Bul+18] Carol J Bult, Judith A Blake, Cynthia L Smith, James A Kadin, et al. "Mouse Genome Database (MGD) 2019". In: *Nucleic Acids Research* 47.D1 (2018), pp. D801–D806. DOI: 10.1093/nar/gky1056.

[CGV18] P. Cinaglia, P. H. Guzzi, and P. Veltri. "INTEGRO: an algorithm for data-integration and disease-gene association". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2018, pp. 2076–2081. DOI: 10.1109/BIBM.2018.8621193.

[Con18] Gene Ontology Consortium. "The Gene Ontology resource: 20 years and still GOing strong". In: *Nucleic Acids Research* 47.D1 (2018), pp. D330–D338. DOI: 10.1093/nar/gky1055.

[Coo+18] Charles E Cook, Rodrigo Lopez, Oana Stroe, Guy Cochrane, et al. "The European Bioinformatics Institute in 2018: tools, infrastructure and training". In: *Nucleic Acids Research* 47.D1 (2018), pp. D15–D22. DOI: 10.1093/nar/gky1124.

[Doğ18] Tunca Doğan. "HPO2GO: prediction of human phenotype ontology term associations for proteins using cross ontology annotation co-occurrences". In: *PeerJ* 6 (2018), e5298. DOI: 10.7717/peerj.5298.

[Dom+18] Daniel Domingo-Fernandez, Charles Tapley Hoyt, Carlos Bobis-Álvarez, Josep Marin-Llao, et al. "ComPath: An ecosystem for exploring, analyzing, and curating mappings across pathway databases". In: *NPJ systems biology and applications* 5.1 (2018), p. 3. DOI: 10.1101/353235.

[Gau+18] Jeff Gauthier, Antony T Vincent, Steve J Charette, and Nicolas Derome. "A brief history of bioinformatics". In: *Briefings in Bioinformatics* 3 (2018). DOI: 10.1093/bib/bby063.

[Gro+18] Christian Grove, Scott Cain, Wen J. Chen, Paul Davis, et al. "Using WormBase: A Genome Biology Resource for Caenorhabditis elegans and Related Nematodes". In: *Eukaryotic Genomic Databases: Methods and Protocols*. Ed. by Martin Kollmar. New York, NY: Springer New York, 2018, pp. 399–470. ISBN: 978-1-4939-7737-6. DOI: 10.1007/978-1-4939-7737-6_14.

[HTK18] Winston A. Haynes, Aurelie Tomczak, and Purvesh Khatri. "Gene annotation bias impedes biomedical research". eng. In: *Scientific Reports* 8.1 (2018), p. 1362. ISSN: 2045-2322. DOI: 10.1038/s41598-018-19333-x.

[HDH18] Charles Tapley Hoyt, Daniel Domingo-Fernandez, and Martin Hofmann-Apitius. "BEL Commons: an environment for exploration and analysis of networks encoded in Biological Expression Language". In: *Database* (2018). DOI: 10.1101/288274.

[MPP18] Noël Malod-Dognin, Julia Petschnigg, and Nataša Pržulj. "Precision medicine–A promising, yet challenging road lies ahead". In: *Current Opinion in Systems Biology* 7 (2018), pp. 1–7. DOI: 10.1016/j.coisb.2017.10.003.

[Mou+18] Fleur Mougin, David Auber, Romain Bourqui, Gayo Diallo, et al. "Visualizing omics and clinical data: Which challenges for dealing with their variety?" In: *Methods* 132 (2018), pp. 3–18. DOI: 10.1016/j.ymeth.2017.08.012.

[Nin+18] Wanshan Ning, Shaofeng Lin, Jiaqi Zhou, Yaping Guo, et al. "WocEA: The visualization of functional enrichment results in word clouds." In: *Journal of genetics and genomics= Yi chuan xue bao* 45.7 (2018), p. 415. DOI: 10.1016/j.jgg.2018.02.008.

[Tom+18] Aurelie Tomczak, Jonathan M. Mortensen, Rainer Winnenburg, Charles Liu, et al. "Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations". In: *Scientific Reports* 8.1 (2018), p. 5115. DOI: 10.1038/s41598-018-23395-2.

[Yu18] Guangchuang Yu. "clusterProfiler: universal enrichment tool for functional and comparative study". In: *bioRxiv* (2018), p. 256784. DOI: 10.1101/256784.

[Zha+18] Jiongmin Zhang, Ke Jia, Jinmeng Jia, and Ying Qian. "An improved approach to infer protein-protein interaction based on a hierarchical vector space model". In: *BMC Bioinformatics* 19.1 (2018), p. 161. DOI: 10.1186/s12859-018-2152-z.

[ZW18] Chenguang Zhao and Zheng Wang. "GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms". In: *Scientific reports* 8.1 (2018), p. 15107. DOI: 10.1038/s41598-018-33219-y.

[Fab+19] Fabio Fabris, Daniel Palmer, João Pedro de Magalhães, and Alex A Freitas. "Comparing enrichment analysis and machine learning for identifying gene properties that discriminate between gene classes". In: *Briefings in bioinformatics* (2019). DOI: 10.1093/bib/bbz028.

[Hoy+19] Charles Tapley Hoyt, Daniel Domingo-Fernández, Sarah Mubeen, Josep Marin Llaó, et al. "Integration of Structured Biological Data Sources using Biological Expression Language". In: *BioRxiv* (2019), p. 631812. DOI: 10.1101/631812.

[Rau+19] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, et al. "g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)". In: *Nucleic Acids Research* (2019). DOI: 10.1093/nar/gkz369.

[Wer+19] Méline Wery, Olivier Dameron, Jacques Nicolas, Elisabeth Remy, et al. "Formalizing and enriching phenotype signatures using Boolean networks". In: *Journal of theoretical biology* 467 (2019), pp. 66–79. DOI: 10.1016/j.jtbi.2019.01.015.

[Woo+19] Valerie Wood, Antonia Lock, Midori A Harris, Kim Rutherford, et al. "Hidden in plain sight: what remains to be discovered in the eukaryotic proteome?" In: *Open biology* 9.2 (2019), p. 180241. DOI: 10.1101/469569.

Appendix A

Additional information for chapter 4

For the used datasets [C-260] and [B-340], we represent the percentage of covered genes provided by DAVID and the different semantic similarity measures used in the workflow developed in section 4.2. figures A.1 and A.2 show several genes sets with low gene coverage due to the fact that, for each gene set, an important number of genes are still unknown.

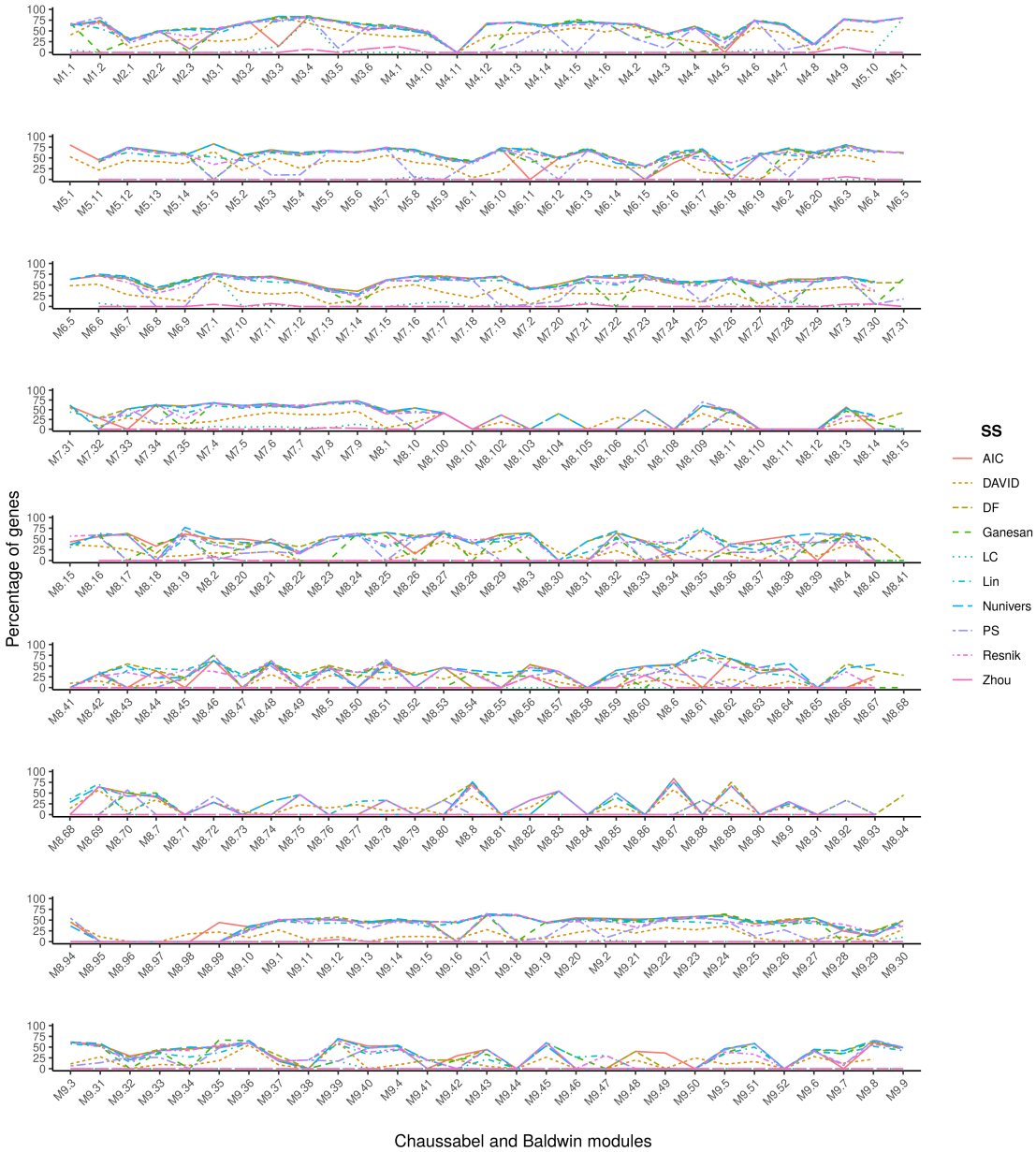


FIGURE A.1: Percentage of covered genes for each gene set in the [C-260] dataset.

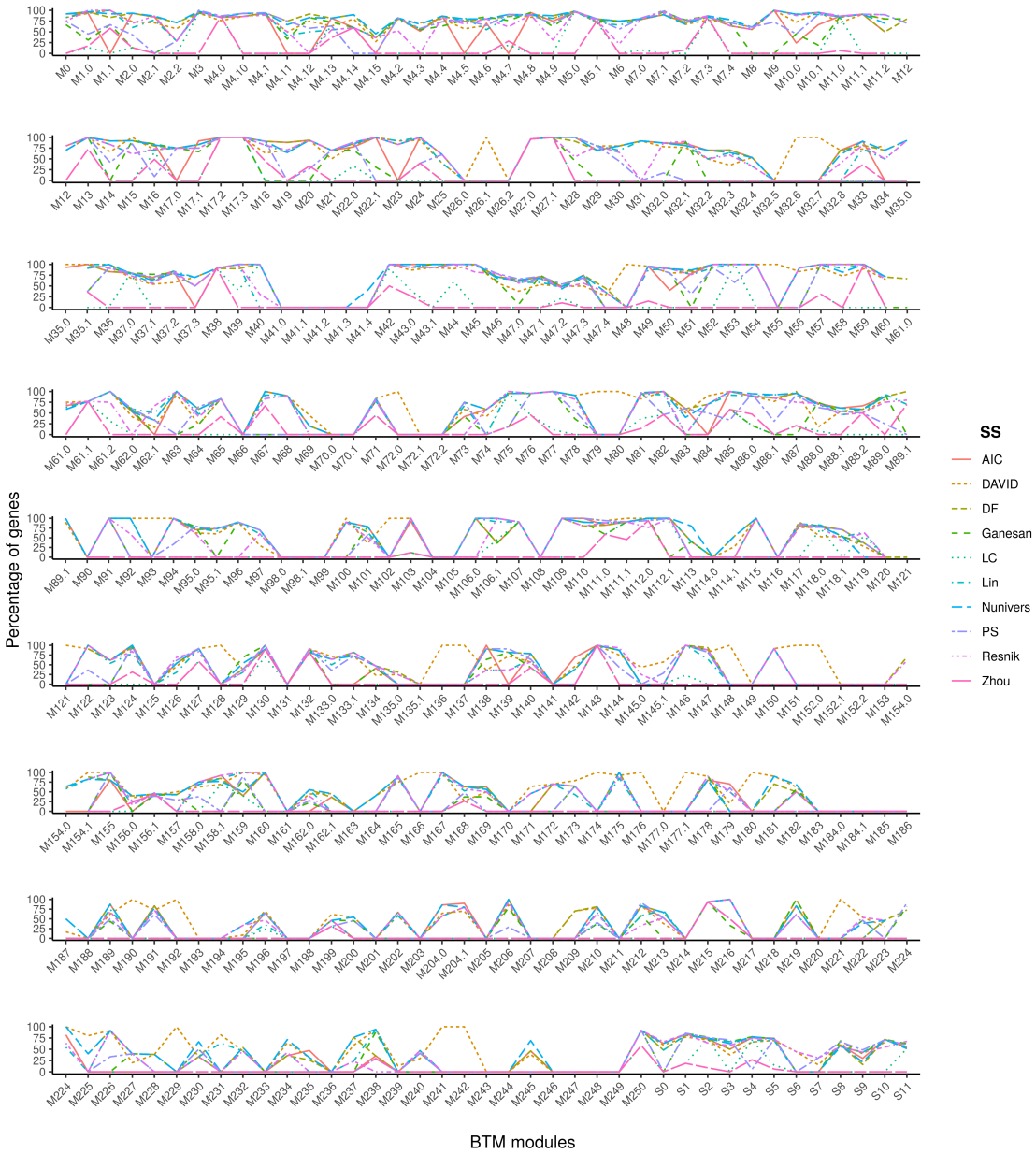


FIGURE A.2: Percentage of covered genes for each gene set in the [B-346] dataset.

Appendix B

Additional information for chapter 5

For each knowledge resource added in GSAn (DO and *Reactome*), we present the *top ten* terms that annotate the highest number of genes.

Disease Ontology

Table B.1 shows the *top ten* DO terms that annotate the highest number of genes. We can observe that most of the DO terms have a depth of 5. This low depth means a low level in detail in the gene annotation. Focusing on the DO term annotating the highest number of genes, *kidney cancer* (DOID:263, with 2,557 annotated genes), we observe that this term has 46 descendant terms. By looking at the direct children (table B.2), we can notice that only three of them annotate genes and the number of annotated genes is insignificant (*e.g.*, *nephroblastoma* annotates 38 out of 2,557 genes). This shows a lack of details in the DO annotation.

DO name	DO id	IC Mazandu	Depth	Number of genes	Number of descendants
kidney cancer	DOID:263	25.86	5.0	2,557	46
liver cancer	DOID:3571	51.41	5.0	758	17
endometrial cancer	DOID:1380	51.11	7.0	742	18
prostate cancer	DOID:10283	36.41	6.0	652	20
epilepsy	DOID:1826	14.66	5.0	586	63
breast cancer	DOID:1612	21.31	5.0	536	80
rheumatoid arthritis	DOID:7148	16.60	7.0	506	1
skin cancer	DOID:4159	23.37	5.0	449	83
lung cancer	DOID:1324	27.04	5.0	411	53
stomach cancer	DOID:10534	26.94	5.0	350	30

TABLE B.1: *Top ten* DO terms annotating the highest number of genes.

DO name	DO id	IC Mazandu	Depth	Number of genes	Number of descendants
nephroblastoma	DOID:2154	28.06	6.0	38	7
renal carcinoma	DOID:4451	28.06	6.0	2	15
kidney sarcoma	DOID:4242	28.06	6.0	1	4
mesoblastic nephroma	DOID:4772	28.06	6.0	0	3
renal pelvis carcinoma	DOID:4919	28.06	6.0	0	6
childhood kidney cancer	DOID:3675	28.06	6.0	0	5
kidney liposarcoma	DOID:5699	57.44	8.0	0	0
kidney hemangiopericytoma	DOID:262	28.06	6.0	0	0
malignant cystic nephroma	DOID:7571	28.06	6.0	0	0

TABLE B.2: Information of the child terms of the *kidney cancer* DO term.

Reactome

Table B.3 shows the *top ten Reactome* terms that annotate the highest number of genes. We can observe that most of the *Reactome* terms have a depth of 4. Contrary to DO, most of *Reactome* terms have no descendant term. Although the depth seems to be low, these terms are specific with a high level of detail.

Reactome name	Reactome id	IC Mazandu	Depth	Number of genes	Number of descendants
Neutrophil degranulation	R-HSA-6798695	9.71	3.0	765	0
Olfactory Signaling Pathway	R-HSA-381753	11.30	5.0	394	0
Antigen processing: Ubiquitination & Proteasome degradation	R-HSA-983168	10.32	4.0	392	0
Generic Transcription Pathway	R-HSA-212436	9.92	3.0	343	65
mRNA Splicing - Major Pathway	R-HSA-72163	10.24	4.0	269	0
Neddylation	R-HSA-8951664	10.93	3.0	235	0
Ub-specific processing proteases	R-HSA-5689880	12.54	4.0	221	0
Rho GTPase cycle	R-HSA-194840	9.51	3.0	221	0
GPCR downstream signalling	R-HSA-388396	9.51	3.0	220	36
G alpha (i) signalling events	R-HSA-418594	11.30	4.0	218	23

TABLE B.3: *Top ten Reactome* terms annotating the highest number of genes.