Méthodes sémantiques pour la comparaison inter-espèces de voies métaboliques : application au métabolisme des lipides chez l'humain, la souris et la poule

Charles Bettembourg

ANNÉE 2013

UNIVERSITÉ DE RENNES 1

THÈSE / UNIVERSITÉ DE RENNES 1

sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Biologie

École doctorale VAS

présentée par

Charles Bettembourg

Préparée au sein des unités de recherche UMR1348 PEGASE et UMR6074 IRISA
PEGASE : Physiologie, Environnement et Génétique pour l'Animal et les Systèmes d'Élevage
IRISA : Institut de recherche en informatique et systèmes aléatoires

---

Méthodes sémantiques pour la comparaison inter-espèces de voies métaboliques : application au métabolisme des lipides chez l'humain, la souris et la poule

Thèse soutenue à Rennes le 16 décembre 2013

devant le jury composé de :

Christine FROIDEVAUX
Professeur, Université Paris-Sud / *rapportrice*

Philippe BESSIÈRES
Directeur de recherche, INRA / *rapporteur*

Nathalie AUSSENAC-GILLES
Directrice de recherche, CNRS / *examinatrice*

Philippe VANDENKOORNHUYSE
Professeur, Université de Rennes 1 / *examinateur*

Christian DIOT
Directeur de recherche, INRA / *directeur de thèse*

Olivier DAMERON
Maître de conférences, Université de Rennes 1
co-directeur de thèse

REMERCIEMENTS

Je tiens tout d'abord à remercier mes deux directeurs de thèse, Olivier Dameron et Christian Diot pour leur encadrement. Malgré un emploi du temps parfois très chargé, ils ont toujours eu à cœur de suivre de près l'évolution de mes travaux. Mon agenda se souvient de plus de 120 réunions au cours desquelles le projet de thèse s'est peu à peu concrétisé, et de plus de 2500 mails échangés entre Olivier, Christian et moi. Sans parler des heures d'échanges par messagerie instantanée pour émettre une idée, la discuter et se tenir informé de son devenir. Merci pour cette implication constante, et merci pour votre confiance.

Je suis aussi reconnaissant envers les rapporteurs de cette thèse, Christine Froidevaux et Philippe Bessières, pour l'évaluation de mon travail et pour l'intérêt qu'ils y ont porté. Merci aussi à Nathalie Aussenac-Gilles et à Philippe Vandenkoornhuyse pour avoir accepté de faire partie de mon jury.

Je remercie les membres de mon comité de thèse, Jacques Mourot, Emmanuelle Becker, Bernard Gibaud, Thomas Faraut et Pierre-Yves Le Bail pour leur suivi et leurs conseils apportés au cours de nos discussions.

Ce travail a été possible grâce à un financement placé sur ce sujet situé à l'interface de la biologie et de l'informatique par le choix du président de l'Université de Rennes 1, Guy Cathelineau que je tiens à remercier.

Je souhaite remercier ensuite les membres des trois équipes qui m'ont accueilli :

Un grand merci à l'équipe Génétique et Génomique de l'unité PEGASE (INRA) qui m'a accueilli pendant presque quatre ans, dès mon stage de master 2. Je pense d'une part à ceux qui continuent à faire tourner le labo : Christian Diot, Pascale Leroy, Sandrine Lagarrigue, Olivier Demeure, Pierre-François Roux, Frédéric Hérault, Olivier Filangi, Colette Désert, Frédéric Lecerf, Sophie Allais, Hélène Romé, Jean-Marc Fraslin et Magalie Houée et d'autre part à ceux qui y sont passés ou en sont parti : Christine Gourbe, Cécile Duby, Walid Bedhiafi, Yoannah François, Thomas Obadia, Aymeric Antoine-Lorquin, Julien Navarro et Émile Richard. Enfin, mille mercis aux anciens doctorants avec qui j'ai passé des moments inoubliables : Yvan Le Bras, Marion Ouédraogo, Yuna Blum et Xiaoqiang Wang. Ils m'avaient mis dans l'ambiance dès mon premier jour au labo, en stage de master 2, avec la préparation et le tournage d'un court-métrage présentant leurs travaux. Comment

mieux commencer ? Mais j'aurai l'occasion de reparler de courts-métrages un peu plus loin !

Je remercie également les membres de l'U936 (INSERM) avec qui j'ai interragi en stage avant la thèse, et pendant les deux premières années de celle-ci : Anita Burgun, Arnaud Rosier, Delphine Rossille, Isabelle Stévant, Nicolas Schnel et Thomas Bernicot.

Depuis 2013 j'ai aussi été accueilli dans le département Data and Knowledge Management à l'IRISA (INRIA) où j'ai pu interragir avec des membres des équipes Dyliss et Genescale et des membres de la plateforme bio-informatique GenOuest : Anne Siegel, Dominique Lavenier, Olivier Collin, Jacques Nicolas, Pierre Peterlongo, Sylvain Prigent, Nicolas Maillet, Guillaume Chapuis, Geoffroy Andrieux, Vincent Picard, Gaëlle Garet, Mathilde Le Boudic-Jamin, Valentin Wucher, Guillaume Collet, Olivier Quenez, Jeanne Cambefort, Anthony Bretaudeau et Olivier Sallou.

Je tiens à remercier ici des personnes d'autres laboratoires qui m'ont elles aussi permis d'avancer dans mes travaux. Merci à Nolwenn Le Meur (EHESP), avec qui nous avons étudié l'évolution de la complexité de Gene Ontology. Merci à Dietrich Rebholz-Schuhmann (Université de Zurich) qui m'a permis de présenter en séminaire mes travaux concernant GO2PUB au groupe Uniprot de l'EBI.

Cette thèse fait suite à des travaux initiés dès mon premier stage de master et j'en profite pour remercier celles et ceux que j'ai rencontré grâce à ce master et avec qui je suis resté en contact. Certains se retrouveront à plusieurs endroit dans cette section de remerciements, ce qui me laisse penser que notre petite communauté "master MSB" s'est formé à l'époque sur des bases solides, préludant à de longues amitiés aussi bien professionnelles que personnelles. Merci donc à Nicolas Maillet, Sylvain Prigent, Geoffroy Andrieux, Thomas Bernicot, Nicolas Schnel, Sylvain Léonard, Thomas Vernet, Tristan Bitard Feildel, Isabelle Stévant, Mathilde Le Boudic-Jamin, Élodie Ruelle, Charlotte Paillette, Thomas Obadia, Arnaud Le Cavorzin, Damien Choisne, Alexandre Cormier, Julien Navarro et Émile Richard.

Je souhaite aussi remercier les membres des associations dans lesquelles j'ai été impliqué. Merci donc à celles et ceux avec qui on a fait vivre DocAIR : Marie Verbanck, Cécile Sauder, Thierry Le Naou, Bertrand Vautier, Marion Ouédraogo, Yuna Blum et Pierre-François Roux. Merci également au bureau de LUCA avec qui nous avons souvent collaboré : Emmanuel Gallaud, Leslie Ratié, Jocelyn Plassais et Nicolas Loyer. Et merci à Joseph Chazalon et à Nicomaque pour avoir réussi à coordonner les différentes associations de doctorants rennaises pour créer des évènements de grande ampleur tels que le forum Docteurs & Entreprises et le festival Sciences en Cour[t]s. Je fermerai d'ailleurs la parenthèse assos en parlant de ce festival grâce auquel nous avons pu faire sortir nos chères thématiques scientifiques de nos labos. Grâce à Sciences en Cour[t]s, j'ai pu participer à la réalisation de quatres films, merci donc à Yvan, Marion, Yuna et Xiao pour *la Poule et la Truite, une fable moderne*, à Geoffroy et Nico pour *Bioinformaticus*, à Marie et Cécile pour *Statistix et le problème de la potion magique* et à Pef et Hélène pour *Quand les poules ont eu des dents*. J'ai également pu participer à l'organisation de l'édition 2013 du festival, pour laquelle je remercie toute l'équipe : Marie Verbanck, Cécile Sauder, Coraline Lafon, Yuna Blum, Sylvain Prigent, Nicolas Maillet, Gaëlle Garet et Sophie Allais.

Enfin, *last but not least*, je tiens à remercier très sincèrement ma famille, qui m'a tou-

jours soutenu. Merci donc en particulier à mon père, qui m'a toujours poussé à donner le meilleur de moi-même, à mon grand-père, qui voit enfin le bout de ces looooongues études et à ma grand-mère qui sait sans doute maintenant qu'hélas non, même en travaillant dans la recherche, je n'ai pas trouvé de remède aux rhumatismes... Un grand merci également à Francine, Alexandra et Fabien pour leur soutien, et enfin merci à Xavier-Alexandre et à Louis-Alexandre pour être les plus merveilleux neveux du monde, petites graines de docteurs !

Méthodes sémantiques pour la comparaison inter-espèces de voies métaboliques : application au métabolisme des lipides chez l'humain, la souris et la poule

La comparaison inter-espèces de voies métaboliques est une problématique importante en biologie. Elle constitue un enjeu aussi bien pour la santé humaine que pour l'agronomie. Actuellement, les connaissances sont générées à partir d'expériences sur un nombre relativement limité d'espèces dites modèles. Mieux connaître une espèce permet de valider ou non une inférence faite à partir de ces données expérimentales. C'est aussi nécessaire pour déterminer si ou dans quelle mesure des résultats obtenus sur une espèce modèle peuvent être transposés à une autre espèce.

Cette thèse propose une méthode de comparaison inter-espèces de voies métaboliques. Cette méthode compare chaque étape d'une voie métabolique en exploitant les annotations dans Gene Ontology qui leur sont associées. Ce travail (i) valide l'intérêt des mesures de similarités sémantiques pour interpréter ces annotations, (ii) propose d'utiliser conjointement une mesure de particularité sémantique et (iii) propose une méthode basée sur des motifs de similarité et de particularité pour interpréter chaque étape de voie métabolique. Les différentes étapes de cette approche sont appliquées à l'étude comparative du métabolisme des lipides chez l'Homme, la souris et la poule.

De nombreux produits de gènes interviennent tout au long d'une voie métabolique. Des annotations peuvent être associées à ces produits de gènes afin de décrire leurs rôles biologiques. En reposant sur une ontologie partagée, ces annotations permettent de comparer les données d'espèces différentes et de tenir compte de différents degrés de précision. Il existe de nombreuses mesures sémantiques qui quantifient la similarité entre des produits de gènes en fonction des annotations qu'ils ont en commun. Nous en avons identifié et utilisé une adaptée à la problématique de comparaison inter-espèces.

En se focalisant sur la part commune aux produits de gènes comparés, les mesures de similarité sémantiques ignorent les caractéristiques spécifiques d'un seul produit de gène. Or la comparaison inter-espèces de voies métaboliques se doit de quantifier non seulement la similarité des produits de gènes qui interviennent dans celles-ci, mais également leurs particularités. Nous avons développé une mesure de particularité sémantique répondant à cette problématique. Pour chaque étape de voie métabolique, nous calculons un profil composé de sa valeur de similarité et de ses deux valeurs de particularité sémantiques.

Concernant l'interprétation des résultats, il n'est pas possible d'établir formellement que deux produits de gènes sont similaires ou que l'un d'eux a des particularités significatives sans disposer d'un seuil de similarité et d'un seuil de particularité. Jusqu'à présent, ces interprétations se faisaient sur la base d'un seuil implicite ou arbitraire. Pour combler ce manque, nous avons développé une méthode de définition de seuils pour les mesures de similarité et de particularité sémantiques.

Nous avons enfin appliqué une mesure de similarité inter-espèces et notre mesure de particularité pour comparer le métabolisme des lipides entre l'Homme, la souris et la poule. Nous avons pu interpréter les résultats à l'aide des seuils que nous avions définis. Chez les trois espèces, des particularités ont pu être observées, y compris au niveau de produits de gènes similaires. Elles concernent notamment des processus biologiques et

des composants cellulaires. Les fonctions moléculaires présentent une forte similarité et peu de particularités. Ces résultats sont biologiquement pertinents.

Semantic methods for the cross-species metabolic pathways comparison : application to human, mice and chicken lipid metabolism

Cross-species comparison of metabolic pathways is an important task in biology. It is a major stake for both human health and agronomy. Currently, knowledge is acquired from some experiments on a relatively low number of species referred to as “models”. A better understanding of a species determines whether to validate or not an inference made from these experimental data. It also determines whether or to what extent results obtained on model species can be transposed to another species.

This thesis proposes a cross-species metabolic pathways comparison method. Our method compares each step of a metabolic pathway using the associated Gene Ontology annotations. This work (i) validates the interest of the semantic similarity measures for interpreting these annotations, (ii) proposes to use jointly a semantic particularity measure and (iii) proposes a method based on similarity and particularity patterns to interpret each metabolic pathway step. We applied the different steps of this approach to the comparative study of lipid metabolism for human, mice and chicken.

Several gene products are involved throughout a metabolic pathway. They are associated to some annotations in order to describe their biological roles. Based on a shared ontology, these annotations allow to compare data from different species and to take into account several level of abstraction. Several semantic measures quantifying the similarity between gene products from their annotations have been developed previously. We have identified and used a semantic similarity measure appropriate for cross-species comparisons.

Because they focus on the common part of the compared gene products, the semantic similarity measures ignore their specific characteristics. Therefore, cross-species metabolic pathways comparison has to quantify not only the similarity of the gene products involved, but also their particularity. We have developed a semantic particularity measure addressing this issue. For each pathway step, we proposed to create a profile combining its semantic similarity and its two semantic particularity values.

Concerning the results interpretation, it is not possible to establish formally that two gene products are similar or that one of them have some significant particularities without having a similarity threshold and a particularity threshold. So far, these interpretations were based on an implicit or an arbitrary threshold. To address this gap, we developed a threshold definition method for the semantic similarity and particularity measures.

We last applied a cross-species similarity measure and our particularity measure to compare the lipid metabolism between human, mice and chicken. We then interpreted the results using the previously defined thresholds. In all three species, we observed some particularities, including on similar genes. They concerned notably some biological processes and cellular components. The molecular functions present a strong similarity and few particularities. These results are biologically relevant.

AVANT-PROPOS

Contexte

Cette thèse a été réalisée sous la direction de Christian Diot et Olivier Dameron, au sein des équipes Génétique et Génomique (UMR PEGASE INRA - Agrocampus Ouest) et Modélisation Conceptuelle des Connaissances Biomédicales (UMR 936 INSERM - université de Rennes 1) puis DYnamics, Logics and Inference for biological Systems and Sequences (UMR IRISA INRIA - CNRS). Le point de départ de ce travail est l'existence de difficultés rencontrées lors de l'étude du métabolisme des lipides chez la poule (*Gallus gallus*). En effet, bien que *Gallus gallus* compte parmi les espèces dites modèles pour l'étude de phénomènes biologiques, les fonctions de la plupart de ses gènes et ses différents métabolismes sont encore mal connus. Cela conduit à un processus de transposition de connaissances relatives à une espèce mieux connue, comme l'Homme (*Homo sapiens*) ou la souris (*Mus musculus*). Or, on ne dispose pas de critères précis pour juger si cette opération est légitime, d'autant plus que *Gallus gallus*, à la différence de *Homo sapiens* et *Mus musculus*, n'est pas un mammifère. Il n'existe pas de méthode formelle de comparaison permettant de déterminer si des différences entre les séquences, entre les annotations de produits de gènes et entre les réactions des voies métaboliques sont ou non associées à des différences de traits phénotypiques observés.

Ces observations ont motivé une collaboration entre les différentes équipes mentionnées précédemment. Christian Diot et l'équipe G&G ont proposé la problématique et fourni l'expertise biologique relative au métabolisme des lipides chez *Gallus gallus*. Olivier Dameron et l'UMR 936, puis Dyliss, ont défini le cadre informatique et sémantique requis pour traiter cette problématique et ont accompagné les développements méthodologiques réalisés. Cette thèse se situe donc à la croisée de la biologie et de l'informatique, avec l'ambition d'apporter des solutions pertinentes à un problème biologique en développant et en utilisant des méthodes et outils sémantiques.

La problématique de comparaison fonctionnelle entre espèces n'est pas spécifique au seul métabolisme des lipides. Les développements proposés ici se veulent avant tout génériques ; ils peuvent être appliqués à n'importe quel métabolisme et à n'importe quelle espèce (sous réserve d'un minimum de connaissance disponible).

Structure du manuscrit

Le premier chapitre du manuscrit expose le contexte biologique de cette thèse et définit notre problématique et notre objectif.

Le deuxième chapitre permet d'identifier les ressources et méthodes pertinentes disponibles et les besoins de nouveaux développements.

Les deux chapitres suivants décrivent les méthodes de comparaisons sémantiques d'annotations de produits de gènes. Ils couvrent respectivement le développement d'une nouvelle mesure de particularité sémantique et l'interprétation conjointe des valeurs de similarité et de particularité sémantiques.

Le chapitre 5 concerne l'application des méthodes précédemment décrites à la comparaison inter-espèces de voies métaboliques. Il se focalisera principalement sur le métabolisme des lipides chez la poule, la souris et l'Homme.

Enfin un dernier chapitre décrit l'apport des approches développées au cours de ce travail de thèse dans d'autres domaines, comme la recherche bibliographique, la comparaison fonctionnelle de gènes dupliqués et l'évolution de Gene Ontology.

TABLE DES MATIÈRES

Première partie

État de l’art

CHAPITRE 1

INTRODUCTION

Dans ce chapitre, nous présentons le contexte biologique, rappelons ce qu'est une voie métabolique, élaborons la problématique et définissons l'objectif de cette thèse. Comme mentionné dans l'avant-propos, le point de départ de ce travail a été un constat de difficultés dans l'étude du métabolisme des lipides chez la poule. Nous expliquerons donc tout d'abord le fonctionnement général du métabolisme des lipides tel qu'on le connaît grâce à l'étude de l'Homme et du modèle murin. Puis nous citerons les particularités connues de ce métabolisme chez les oiseaux, notamment chez la poule. Nous verrons que les connaissances concernant la structure des voies métaboliques reflètent parfois mal ces particularités. Cela nous conduira à identifier le besoin d'une nouvelle approche systématique prenant en compte non seulement les données relatives à la structure des voies métaboliques que l'on souhaite étudier, mais également les connaissances disponibles sur les gènes qui interviennent dans ces voies métaboliques.

Sommaire

1 Contexte biologique

Une voie métabolique est une suite de réactions biochimiques intervenant dans un organisme afin d'en assurer le bon fonctionnement. Les lipides constituent avec les protides et les glucides une des trois classes de nutriments énergétiques, indispensables à la vie. Les lipides sont des molécules hydrophobes ou amphipathiques [1] issues pour tout ou partie de la condensation de thio-esters (acides gras, glycérolipides, glycérophospholipides, sphingolipides, glycolipides et polycétides) et/ou de la condensation d'unités isoprènes (prenols et stérols) [Fahy et al., 2009]. Leurs rôles sont multiples. Il permettent la couverture des besoins énergétiques, participent à la constitution des structures micro et macroscopiques de l'organisme et interviennent dans de nombreux mécanismes biochimiques indispensables à la vie. L'étude du métabolisme des lipides chez les oiseaux est importante tant d'un point de vue appliqué, l'engraissement impacte la valeur économique des produits avicoles, que cognitif, au regard de son évolution chez les vertébrés par exemple. De fait, le schéma global du métabolisme lipidique diffère entre les oiseaux et les mammifères. Après avoir présenté le métabolisme des lipides tel qu'on le connaît chez les mammifères, nous aborderons les particularités relevées chez les oiseaux.

1.1 Généralités sur le métabolisme des lipides

Même si cela n'est pas toujours précisé, il convient d'indiquer que les connaissances acquises sur le métabolisme des lipides proviennent majoritairement d'études réalisées sur les espèces modèles, essentiellement mammifères. On verra par la suite que cette présentation générale a trop souvent tendance à occulter les particularités d'autres espèces, plus ou moins éloignées des mammifères au regard de l'évolution.

Les lipides présents dans l'organisme peuvent provenir de l'alimentation ou être néosynthétisés. Les lipides provenant de l'alimentation sont absorbés au niveau de l'intestin grêle. Il s'agit essentiellement de triglycérides, molécules constituées d'un squelette de glycérol dont les trois groupements hydroxyles ont été estérifiés par des acides gras. L'alimentation apporte également des esters de cholestérol, des phospholipides et des

1. Une molécule amphipathique possède à la fois un groupement hydrophile et un groupement hydrophobe.

vitamines liposolubles (A, D, E et K). Ces lipides suivent un trajet bien défini dans l'organisme : à partir de l'intestin grêle, ils vont passer dans le système lymphatique puis dans le sang, où ils seront distribués aux tissus qui en ont besoin. A l'issu de ce circuit, les lipides résiduels seront captés par le foie. Dans l'intestin grêle, les acides et sels biliaires émulsionnent les gouttelettes de lipides alimentaires. La lipase pancréatique libère des monoglycérides et des acides gras qui forment des micelles. La cholestérol-estérase et la phospholipase hydrolysent réciproquement les esters de cholestérols et les phospholipides et libèrent ainsi des acides gras qui entrent aussi dans la composition des micelles. Celles-ci sont absorbées de façon passive par les cellules de l'épithélium intestinal (entérocytes). Les entérocytes absorbent aussi le cholestérol libre et les vitamines liposolubles. Les acides gras libres composés de moins de 12 carbones diffusent directement depuis l'entérocyte vers le foie via le système porte. Dans les entérocytes se produit la re-synthèse des triglycérides, des esters de cholestérol et des phospholipides, qui sont exportés avec les vitamines liposolubles dans le système lymphatique sous forme de chylomicrons [Hussain et al., 1996]. Les chylomicrons sont des sphères de 75 à 1200 nm contenant en leur centre la partie hydrophobe (triglycérides, partie hydrophobe des phospholipides et du cholestérol) des lipides digérés et en périphérie leur partie hydrophile (tête polaire des phospholipides, groupe hydroxyle du cholestérol) ansi que des lipoprotéines (apolipoprotéines). L'apolipoprotéine principale de ces chylomicrons dits naissants est l'apolipoprotéine B-48 (APOB48). Les apolipoprotéines A-I, A-II et A-IV participent également à la composition de ces chylomicrons naissants. Les chylomicrons exportés dans le système lymphatique rejoignent la circulation sanguine au niveau de la veine sous-clavière gauche. Dans la circulation sanguine, ils deviennent matures en acquérant des apolipoprotéines C-II, C-III et E grâce à un échange avec des particules lipidiques de haute densité (HDL). L'apolipoprotéine C-II étant le cofacteur de la lipoprotéine-lipase, celle-ci peut libérer les acides gras contenus dans les chylomicrons afin qu'ils soient absorbés par les cellules des tissus vascularisés. Les chylomicrons transfèrent ensuite leurs apolipoprotéines A-I, A-IV, C-II et C-III aux HDL pour devenir des chylomicrons remnants de 30 à 50 nm de diamètre qui seront reconnus et absorbés par le foie grâce à leurs APOE et APOB-48. La figure 1 résume ce transport des lipides alimentaires.

La lipogenèse (synthèse *de novo* des acides gras) a lieu dans deux tissus distincts : le foie et le tissu adipeux [Bergen et Mersmann, 2005]. Elle assure la synthèse d'acides gras à longue chaîne hydro-carbonée qui seront incorporés dans des triglycérides. Les enzymes clés de cette synthèse sont l'acétyl-CoA carboxylase (ACC, EC $6.4.1.2$), la malate déshydrogénase (EC $1.1.1.39$), et l'acide gras synthase (Fatty Acid Synthase, FAS, EC $2.3.1.85$). Ces enzymes sont stimulées par l'insuline et inhibés par le glucagon. Le précurseur de la lipogenèse est l'acétyl-CoA, qui peut être obtenu à l'issue de la glycolyse (dégradation du glucose), de la β-oxydation des acides gras (principale voie de dégradation des acide gras) ou encore de la dégradation des acides aminés cétogènes. L'acétyl-CoA est produit dans la mitochondrie puis exporté dans le cytoplasme où l'acétyl-CoA carboxylase permet la synthèse du malonyl-CoA. La FAS est un complexe enzymatique permettant la condensation successive d'unités malonyl-CoA sur de l'acétyl-CoA jusqu'à obtention de l'acide palmitique. La figure 2 présente la suite de réactions mises en œuvre pour obtenir une molécule d'acide palmitique.

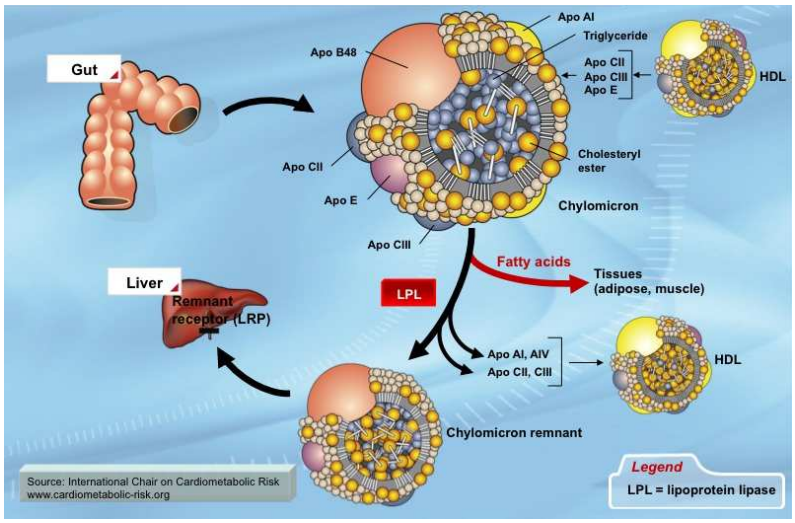


FIGURE 1 – Métabolisme des chylomicrons. Les lipides issus de l'alimentation sont incorporés sous forme de triglycérides et d'esters de cholestérol aux chylomicrons au niveau des intestins. Les chylomicrons vont les distribuer aux tissus vascularisés lors d'un circuit qui les mènera finalement vers le foie.

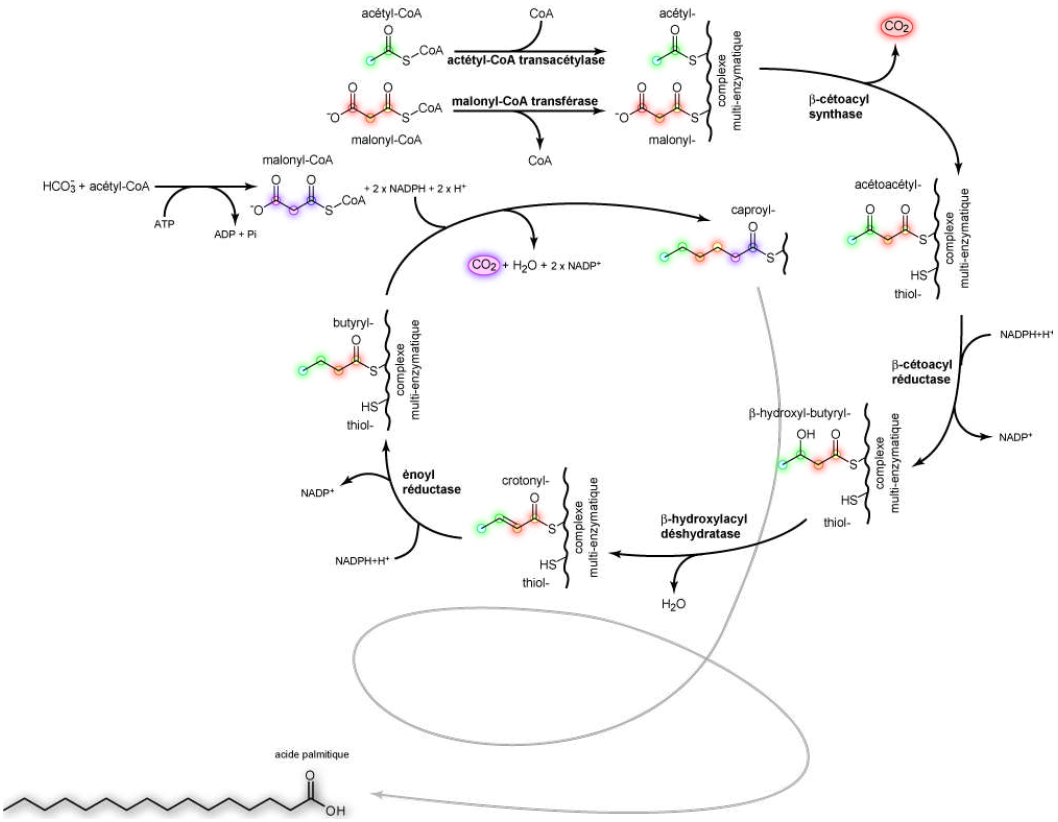


FIGURE 2 – Synthèse de l'acide palmitique. Les acides gras se forment par condensation successives de molécules de malonyl-CoA.

L'acide palmitique sert de base à la construction d'acides gras insaturés à longue chaîne. Ceux-ci sont obtenus par une succession d'élongations (ajout de 2 carbones) et de désaturations (création d'une double liaison). Ce processus a lieu dans le réticulum endoplasmique. La désaturation est assurée par une désaturase capable de catalyser le départ de deux atomes d'hydrogène de la molécule d'acide gras, créant une double liaison carbone/carbone. La position de la double liaison est à la base des deux nomenclatures des acides gras insaturés. Les positions dans les molécules sont définies par rapport au groupement le plus réactif, en l'occurrence le groupement carboxyle pour les acides gras. Ainsi, la Δ9 désaturase crée une double liaison sur l'acide palmitique après le 9$^{\text{ème}}$ carbone depuis le groupe carboxyle pour donner l'acide palmitoléique. Cet acide gras est symbolisés ainsi : (16:1)Δ9. Cependant, la numérotation des carbones dans un acide gras se fait usuellement dans l'autre sens. On décrit ainsi l'appartenance à une « série omega » en comptant la position de la double liaison à partir du groupe méthyl terminal. Ainsi, l'acide palmitoléique (16:1)Δ9 est un acide gras de la série des ω7 (ou n-7). L'Homme a quatre désaturases différentes : Δ9, Δ6, Δ5 et Δ4. N'ayant pas de Δ12 ni de Δ15 désaturase qui n'existent que dans le règne végétal, l'Homme est incapable de synthétiser certains acides gras poly-insaturés, tels que l'acide linoléique (18:2)Δ9,12 et l'acide α-linolénique, (C18:3)Δ9,12,15. Ils sont respectivement précurseur des séries ω6 et ω3, à la base de la synthèse de nombreuses molécules comme des prostaglandines ou l'acide arachidonique. Ces acides gras sont dits essentiels et doivent être apportés par l'alimentation. La figure 3 résume la synthèse des différents acides gras insaturés.

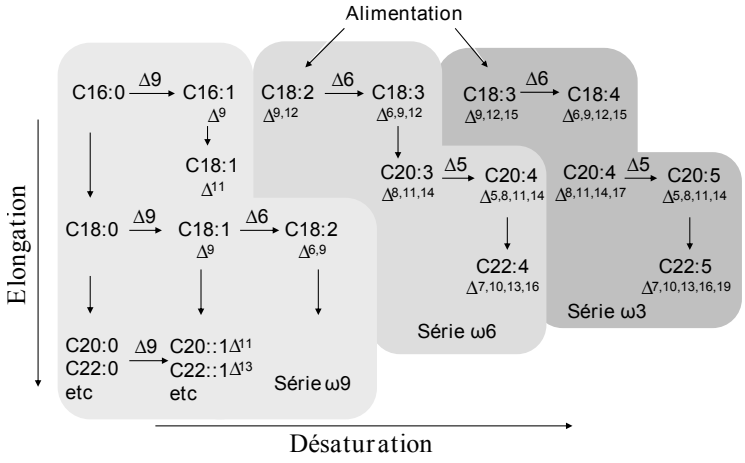


FIGURE 3 – Synthèse des acides gras insaturés. Deux réactions sont répétées pour obtenir des acides gras insaturés de différentes séries : une désaturation puis une élongation. Les précurseurs des acides gras des séries ω6 et ω3 doivent être fournis par l'alimentation.

Les acides gras obtenus lors de la lipogenèse peuvent servir de précurseurs de diverses molécules indispensables au fonctionnement de l'organisme. Ils peuvent également être stockés sous forme de triglycérides par une triple estérification d'une molécule de glycérol. Cette réaction utilise une molécule de Glycérol-3-Phosphate (G3P) dont les fonctions alcool primaire et secondaire sont d'abord estérifiées par deux acides gras pour obtenir un diacylglycérol. Le groupement phosphate du G3P estérifié

est ensuite hydrolysé la phosphatidate phosphatase, ce qui permet l'estérification d'un troisième acide gras. À la place de ce troisième acide gras peut venir s'estérifier un alcool phosphorylé pour donner un phospholipide.

Les mammifères sont également capables de synthétiser du cholestérol. Cette synthèse se fait dans le cytoplasme des cellules du foie et de l'intestin à partir de l'hydroxy-méthyl-glutaryl-CoA (HMG-CoA). Cet HMG-CoA est issu de la condensation de 3 molécules d'acétyl-CoA. L'HMG-CoA réductase transforme l'HMG-CoA en mévalonate. Le mévalonate est précurseur d'isoprénoïdes qui se condensent en squalène, dont les insaturations permettent de former les cycles qui constituent le cholestérol.

Les triglycérides servent de lipides de stockage dans les adipocytes. Ils peuvent être hydrolysés en acides gras par des lipases lors de la lipolyse et libérés dans le sang afin de fournir de l'énergie aux cellules de l'organisme. La lipolyse est activée par les catécholamines (adrénaline et noradrénaline). Les adipocytes jouent également un rôle important dans le phénomène de satiété en étant notamment le siège de la synthèse de la leptine, qui régule l'appétit au niveau de l'hypothalamus.

Les lipides, molécules hydrophobes ou amphipathiques, doivent circuler dans le sang afin d'atteindre leur lieu de stockage ou d'utilisation. C'est l'objet du métabolisme des lipoprotéines. Nous avons vu le transport des lipides alimentaires par les chylomicrons au début de cette section. Les lipides néo-synthétisés sont transportés par des mécanismes similaires utilisant des lipoprotéines. En période post-prandiale, le foie synthétise des lipoprotéines de très faible densité, les VLDL. Elles contiennent des triglycérides, des esters de cholestérol et des apolipoprotéines B-100 et A-I. Comme les chylomicrons, les VLDL doivent apporter les triglycérides aux tissus périphériques. Elles doivent donc obtenir des apolipoprotéines E et des apolipoprotéines C-II afin d'être reconnues et hydrolysées par la lipoprotéine lipase au niveau des cellules périphériques. Comme les chylomicrons, les VLDL obtiennent ces apolipoprotéines par un échange avec des lipoprotéines circulantes de haute densité, les HDL. Déchargées d'une partie de leurs triglycérides, les VLDL diminuent de taille tout en devenant plus denses, elles évoluent en lipoprotéines de densité intermédiaire ou IDL. En raison de la taille réduite des IDL par rapport aux VLDL, les apolipoprotéines C-II perdent leur affinité avec la particule et sont transférées aux VLDL, aux HDL et aux chylomicrons. Il y a également un transfert de triglycérides et de phospholipides des IDL vers les HDL, et d'esters de cholestérol des HDL vers les IDL. Ces derniers échanges conduisent à la dernière étape de l'évolution de ces lipoprotéines qui deviennent des lipoprotéines de faible densité ou LDL. Elles contiennent essentiellement des esters de cholestérol et une apolipoprotéine B-100. Elles peuvent déposer du cholestérol à la surface des membranes des cellules périphériques. Grâce à leur APO B-100 elles sont reconnues par les cellules périphériques, qui les internalisent par endocytose et les hydrolysent totalement.

Les chylomicrons comme les VLDL interragissent avec ce qu'on a appelé des lipoprotéines de haute densité, les HDL. Ces HDL constituent la dernière classe de lipoprotéines. Elles sont synthétisées par le foie et excrétées dans la circulation sanguine. Elles sont constituées essentiellement de phospholipides et d'apolipoprotéines E, A et C, dont elles sont un réservoir circulant pour les autres classes de lipoprotéines. Elles ont également pour rôle la récupération du cholestérol libre déposé à la surface de la mem-

brane des cellules périphériques. Elles piègent ce cholestérol en l'estérifiant, ce qui le retire de la circulation. Le foie retire de la circulation les HDL ayant rempli leur mission ; il les internalise et hydrolyse leurs esters de cholestérol qui entreront dans la compositions de nouvelles lipoprotéines.

Le circuit du cholestérol et des triglycérides est présenté dans les figures 4 et 5.

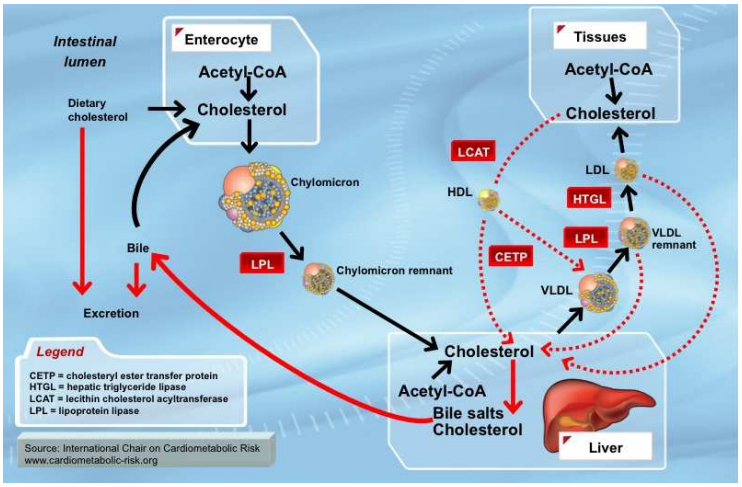


FIGURE 4 – Transport du cholestérol dans le sang. Outre les chylomicrons présentés dans la Figure 1, les VLDL et les HDL participent au transport du cholestérol. Les premiers transportent le cholestérol vers les tissus tandis que les deuxièmes ramènent le cholestérol déposé en excès vers le foie.

Le mauvais fonctionnement du métabolisme des lipides peut être à l'origine de plusieurs pathologies. Comme décrit ci-dessus, les lipides circulent dans le sang ; on parle de lipides plasmatiques. Un excès d'apports en lipides peut causer un dérèglement du taux de ces lipides plasmatiques, et être responsable d'athérosclérose et des pathologies vasculaires associées [Barton, 2013]. La moitié des décès causés par une cardiopathie coronarienne seraient imputables à des taux de cholestérol trop élevés [Stamler et al., 1986; Magnus et Beaglehole, 2001]. Les dérèglements du métabolisme des lipides peuvent avoir des origines génétiques. Ainsi, 15% des cas d'infarctus du myocarde précoces pourraient résulter de troubles héréditaires du métabolisme des lipides [Gaddi et al., 2007].

En dehors de la circulation sanguine, un autre organe pour être fortement impacté en cas de dérèglement du métabolisme des lipides : le foie. La stéatose hépatique correspond à l'infiltration de lipides dans les cellules du parenchyme hépatique. La forme non-alcoolique concerne entre 6% et 24% de la population (un adulte sur trois et un enfant ou un adolescent sur dix aux États-Unis) [Clark et Diehl, 2003; Clark, 2006; Angulo, 2007]. La prévalence est cependant nettement plus importante en cas de surpoids ou d'obésité [Clark, 2006; Angulo, 2007; Papandreou et al., 2007; Moore, 2010]. La stéatose hépatique peut déboucher sur une stéatohépatite, une fibrose voire une cirrhoses du foie ou un carcinome hépatocellulaire [Clark et Diehl, 2003; Qian et Fan, 2005; Reddy et Rao, 2006; Moore, 2010]. La stéatose hépatique non-alcoolique est fortement associée à l'obésité, à la résistance à l'insuline (y compris en raison de diabète), ainsi qu'à un taux

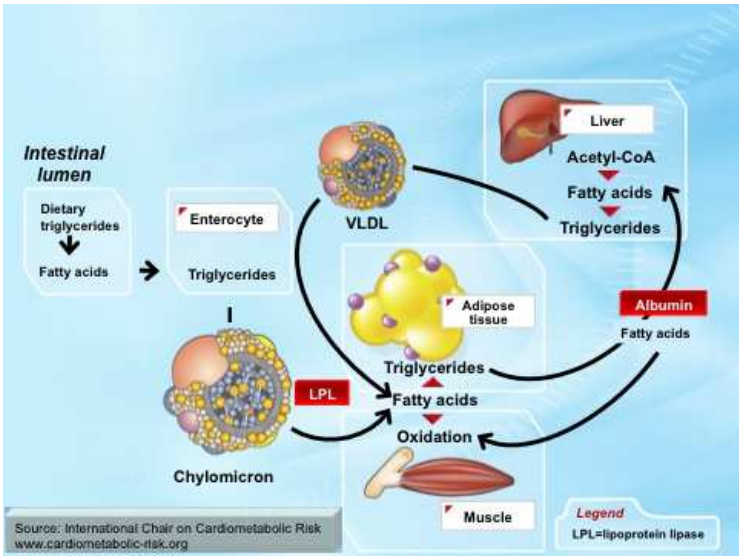


FIGURE 5 – Transport des triglycérides dans le sang. Les chylomicrons transportent les triglycérides provenant de l'alimentation tandis que les VLDL les transportent depuis le foie vers les tissus qui en ont besoin.

élevé de triglycérides ou à un taux faible de lipoprotéines à faible densité [Clark, 2006; Reddy et Rao, 2006].

1.2 Particularités du métabolisme des lipides chez les oiseaux

Nous avons vu les grandes lignes du métabolisme des lipides tracées à partir de connaissances obtenues essentiellement chez les mammifères. D'autres espèces, les oiseaux notamment et le poulet en particulier, présentent cependant des différences par rapport au schéma que nous venons de décrire.

Les oiseaux n'ont pas de vaisseaux lymphatiques intestinaux. Après leur absorption dans l'intestin grêle, les lipides alimentaires sont assemblés dans les entérocytes sous forme de portomicrons (équivalents aux chylomicrons des mammifères) et libérés dans la circulation porte. Les portomicrons vont donc être captés en partie par le foie avant de rejoindre la circulation générale [Fraser et al., 1986].

La lipogenèse est très limitée dans les tissus adipeux ; elle a principalement lieu dans le foie [Hermier, 1997]. Le stockage des triglycérides dépend du substrat lipidique plasmatique issu de l'alimentation et de la synthèse hépatique. L'accumulation excessive et non valorisable de lipides dans les tissus adipeux des poulets de chair est actuellement un problème majeur pour les producteurs [Bourneuf et al., 2006; Daval et al., 2000]. Dans les jeunes poulets de chair approchant leur poids commercial, entre 80 et 85% des acides gras accumulés dans les tissus adipeux sont dérivés de lipides plasmatiques [Griffin et al., 1992]. L'alimentation de ces poulets est pauvre en graisses (moins de 10%) constituées principalement de triglycérides.

Tous les autres triglycérides sont synthétisés dans le foie, dépendant comme chez les mammifères de la disponibilité de glucose alimentaire qui permet d'obtenir de l'acétyl-CoA [Bergen et Mersmann, 2005]. Les triglycérides ne sont pas les seuls lipides à être synthétisés dans le foie, qui est aussi le principal site de synthèse du cholestérol et des phospholipides. Ces lipides, associés à des apolipoprotéines, sont les principaux constituants des lipoprotéines [Hermier, 1997].

Les deux principales classes de particules lipoprotéiques (HDL et VLDL) sont synthétisées et sécrétées par le foie, à destination des tissus de stockage lipidique. Leur partie protéique (apolipoprotéines) y est aussi synthétisée. L'apolipoprotéine B (APOB) et l'apolipoprotéine A-1 (APOA1) sont les deux principales apolipoprotéines chez le poulet [Brown et Dower, 1990]. A la différence des mammifères, la poule n'a pas d'apolipoprotéine E (APOE), mais sa fonction est portée par APOA1 [Daval et al., 2000]. Les triglycérides, le cholestérol, les phospholipides et APOB sont assemblés en VLDL secrétés dans la circulation sanguine. Il en va de même pour la formation des HDL avec APOA1. Les triglycérides s'associent préférentiellement avec APOB pour former des VLDL tandis que les phospholipides et le cholestérol s'associent plutôt avec APOA1 pour former des HDL [Hermier, 1997]. Chez la poule, les triglycérides sont stockées principalement dans les tissus périphériques abdominaux. A la différence des mammifères, ces tissus adipeux ne secrètent pas de leptine, l'hormone de satiété, qui n'existe pas chez la poule [Pitel et al., 2010].

Le transfert des triglycérides depuis les VLDL et les portomicrons dans les tissus adipeux implique leur catabolisme par la lipoprotéine lipase (LPL). La LPL est synthétisée dans les tissus adipeux, les muscles et autres types cellulaires, mais seules les LPL sécrétées et captées à la surface des capillaires sont actives Hermier [1997]. La LPL est l'enzyme dont le taux est limitant pour l'hydrolyse des lipoprotéines plasmatiques riches en triglycérides. L'activité LPL diminue avec une nutrition riche en acides gras insaturés des séries ω3 et ω6.

Un oiseau dont la lipogenèse excède la capacité de synthèse et de sécrétion hépatique de lipoprotéines développe un foie gras. Dans le cas des poules pondeuses, chez lesquelles la stimulation de la lipogenèse par les estrogènes peut conduire au dépassement de la capacité de sécrétion des VLDL, cela peut provoquer une maladie métabolique : le syndrome de foie gras hémorragique, qui réduit la ponte et augmente la mortalité [Hansen et Walzem, 1993]. Les palmipèdes sauvages subissent un engraissement général avant leur migration, leur foie gras servant d'organe de stockage d'énergie. Cette capacité naturelle est utilisée pour la production de foie gras par gavage avec un régime alimentaire riche en glucides. Dans ces conditions, la lipogenèse hépatique augmente radicalement, et le poids du foie peut passer de 100 g à 1 kg en 2 semaines. La stéatose hépatique est due à une accumulation de triglycérides dans les cellules du parenchyme hépatique. Chez l'oie, cela provoque une importante augmentation des concentrations de HDL et VLDL. En outre, ces VLDL contiennent moins de triglycérides, témoignant d'un défaut d'incorporation des triglycérides dans les VLDL, à l'origine de leur accumulation dans le foie chez ces espèces. Chez les poulets, une grande quantité de triglycérides est stockée temporairement dans le foie, mais nécessite ensuite une hydrolyse et une ré-estérification avant d'être sécrétée. Chez les palmipèdes gavés, la régulation hormonale ne permet pas au

foie d'évacuer cet excès de lipides, qui s'accumule [Hermier, 1997].

On le voit, ces quelques exemples suffisent à illustrer des différences qui existent entre un oiseau (la poule) et un mammifère. Ils soulèvent aussi la question de l'analyse des ressemblances et différences dans un cadre plus global.

2 Comparaison inter-espèces : de l'approche structurelle à l'approche fonctionnelle

L'intégralité des réactions biochimiques qui ont lieu dans un organisme sont liées, comme le montre la figure 6 issue de la base de données KEGG. Il est cependant possible de considérer des segments de suites de réactions, qui constituent une voie métabolique. Ces différentes voies métaboliques sont symbolisées par les différentes couleurs de la figure 6.

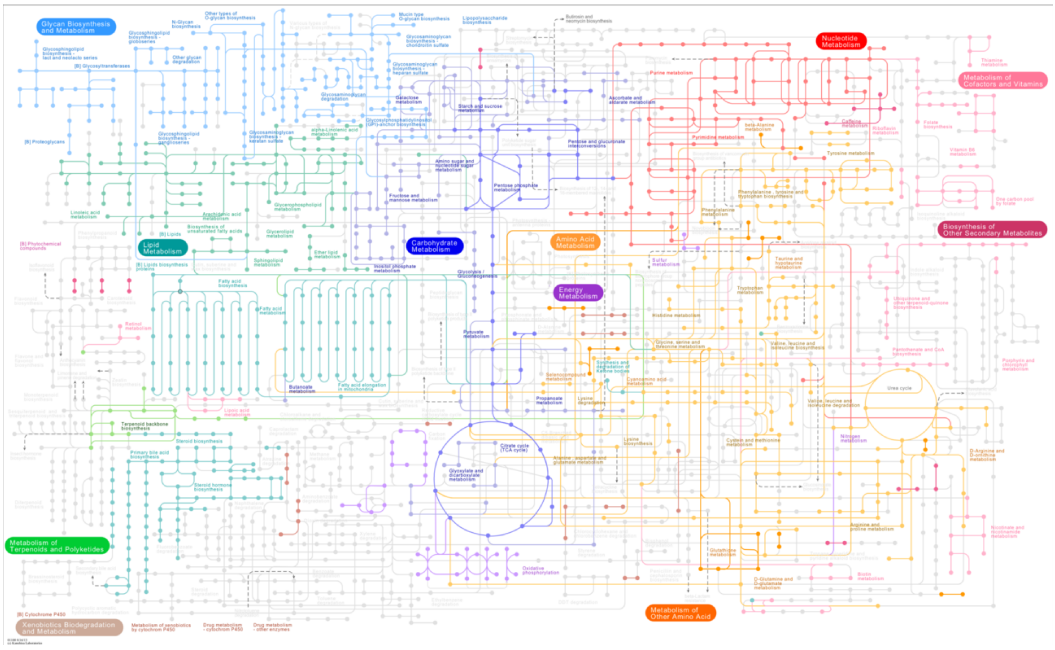


FIGURE 6 – Carte du métabolisme de l'Humain proposée par la base de données KEGG.

Entre deux espèces, une voie métabolique peut être parfaitement identique, différer par quelques réactions chimiques, voire être présente chez une espèce et absente chez une autre. Ainsi, si on considère *Homo sapiens* et *Gallus gallus*, la synthèse de l'acide palmitique se déroule de la même façon, alors le phénomène de satiété fait intervenir des agents différents (absence de leptine chez *Gallus gallus*) et que la lactation est totalement absente chez *Gallus gallus*. La conservation de voies métaboliques entre espèces est liée à leur proximité taxonomique. Il est possible d'évaluer la similarité d'une voie métabolique

analogue entre deux espèces en comparant les réactions présentes chez chacune des espèces.

L'enchaînement des réactions au sein des voies métaboliques des espèces proches, comme les vertébrés, sont souvent rigoureusement identiques. Cela signifie qu'une voie métabolique identique ou très similaire entre deux espèces au niveau de sa structure peut être finalement assez différente au niveau des fonctions biologiques qui dépendent d'elle. On peut ainsi parler de voies métaboliques structurellement identiques ou similaires mais fonctionnellement différentes. On peut également envisager le cas inverse de voies métaboliques dont la structure est différente mais dont les fonctions sont similaires.

Il faut étudier plus en détail les intervenants des réactions pour mieux comprendre ce qui provoque les différences constatées entre espèces. Les réactions des voies métaboliques sont généralement catalysées par des enzymes. Lorsqu'une même réaction est présente chez deux espèces, l'enzyme impliquée peut être codée par un gène homologue. On parle d'homologie quand un gène existe en plusieurs versions dérivant d'une même version originelle à travers un processus d'évolution. Si ces différentes versions appartiennent à des espèces différentes, on parle d'orthologie. Si ces versions co-existent au sein d'une même espèces, on parle de paralogie. Il est également possible qu'une enzyme qui catalyse une même réaction chez deux espèces ne soit pas le produit de l'évolution d'un même gène originel. On parle alors de gènes ayant des fonctions analogues, mais n'ayant aucun lien dans l'évolution.

L'étude des fonctions des gènes a permis d'annoter fonctionnellement ceux-ci, c'est-à-dire d'associer à chaque gène des mots-clés résumant leur fonction. Le vocabulaire employé lors de ce processus d'annotation est formalisé au sein d'une structure appelée Gene Ontology présentée dans le chapitre suivant.

3 Objectif

L'objectif de cette thèse était de développer une méthode et des outils associés pour comparer fonctionnellement les voies métaboliques entre espèces sur la base des annotations des produits de gènes qui y interviennent et en exploitant les connaissances du domaine afin d'interpréter ces annotations. Pour chaque voie métabolique connue, cette méthode avait pour but de vérifier l'identité, ou à défaut, le degré de similarité structurel de cette voie entre plusieurs espèces. Ensuite, la méthode devait être capable d'identifier le degré de similarité et les particularités de chaque produit de gène orthologue intervenant dans chaque voie métabolique chez les espèces d'intérêt. Enfin, la mise en parallèle de la structure d'une voie métabolique et des résultats de la comparaison des gènes qui y interviennent devait permettre de mieux comprendre les différences entre espèces. Ces travaux avaient pour but de confirmer ou d'infirmer la possibilité de prendre en compte des résultats acquis chez une espèce dans l'étude d'une autre.

CHAPITRE 2

MATÉRIEL ET MÉTHODES

Dans ce chapitre, nous présentons les données et les méthodes disponibles pour la comparaison inter-espèces de voies métaboliques. Cette thèse se base sur l'analyse de connaissances existantes pour une meilleure compréhension de phénomènes biologiques. Il n'y a donc pas eu de génération de données expérimentales au cours de ce travail. Cela a demandé une étude des ressources et approches existantes afin de s'assurer de leur pertinence et de leur qualité. Les voies métaboliques sont décrites dans plusieurs grandes bases de données. Les gènes qui y interviennent sont annotés par des ensembles de termes organisés au sein d'une structure sémantique particulière appelée « ontologie ». Des méthodes propres aux ontologies ont été développées par de nombreuses équipes afin de comparer ces ensembles d'annotations. Nous répertorions donc ici les bases de données de voies métaboliques, décrivons les propriétés des ontologies en général et de Gene Ontology en particulier, puis présentons les méthodes de mesure de similarité sémantique qui permettent la comparaison de produits de gènes.

Sommaire

1 Ressources disponibles

1.1 Bases de données de voies métaboliques

Il existe plusieurs bases de données de voies métaboliques. Elles diffèrent par trois aspects principaux. Premièrement, elles peuvent être dédiées à une seule espèce ou à plusieurs. Deuxièmement, chacune d'entre elles définit différemment le découpage des suites de réactions qui constituent une voie métabolique. Troisièmement, le formalisme employé par chaque base de données lui est généralement propre, ce qui rend difficile la comparaison ou la combinaison des données issues de plusieurs bases. Une étude récente a montré que les données disponibles dans les grandes bases de données de voies métaboliques ont un faible niveau de cohérence, d'exhaustivité et de compatibilité [Soh et al., 2010].

1.1.1 Reactome

Reactome [1] est une base de données de voies métaboliques multi-espèces [Croft et al., 2011]. Cependant, le cœur de Reactome concerne l'Humain, les évènements orthologues concernant une vingtaine d'autres espèces étant manuellement inferrés. Les

1. http://www.reactome.org

données sont toutes revues manuellement par des experts biologistes. L'unité de base employée pour décrire une voie métabolique est la réaction. Les différentes entités biologiques participant aux réactions biochimiques forment un réseau d'interactions biologiques et sont groupés au sein de grandes voies métaboliques. Tout le contenu de Reactome est librement disponible dans des formats d'échange standards tels que SBML et BioPAX. SBML encode au format XML des modèles constitués d'entités (molécules) interragissant dans des processus (réactions). BioPAX (Biological Pathway Exchange) est un format standard basé sur RDF/OWL qui a pour but de représenter les voies métaboliques au niveau moléculaire et cellulaire. BioPAX est plus complet que SBML grâce au niveau sémantique apporté par OWL (*Web Ontology Language*), qui permet l'application de raisonnements à l'aide d'outils comme Protégé. La figure 1 présente le nombre de voies métaboliques, réactions, complexes et protéines recensés par Reactome en juin 2013. Grâce à son formalisme standard, sa gratuité et la présence de la poule parmi les organismes disponibles, Reactome a été la base de données de référence pour les travaux menés au cours de cette thèse.

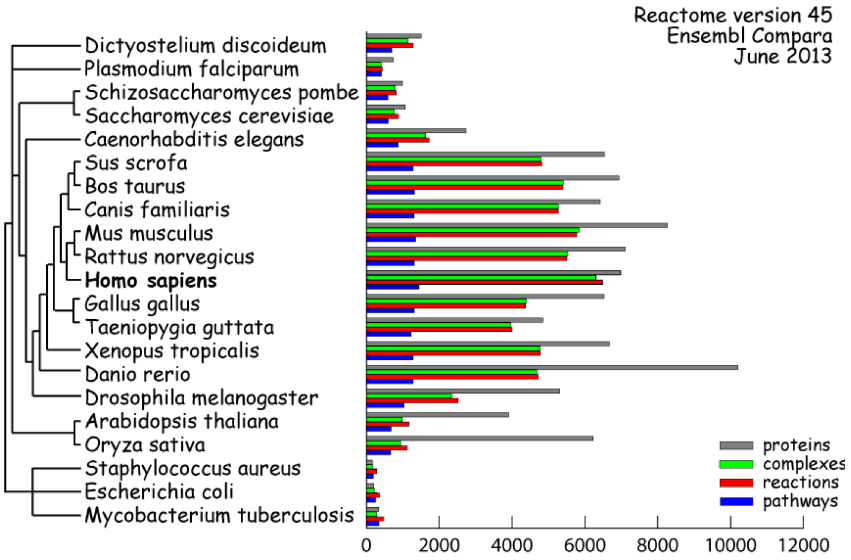


FIGURE 1 – Nombre de voies métaboliques, réactions, complexes et protéines recensés par Reactome en juin 2013.

1.1.2 BioCyc et MetaCyc

BioCyc rassemble près de 3000 bases de données de voies métaboliques, chacune d'entre elles étant mono-espèce, à l'exception d'une seule (MetaCyc) [Caspi et al., 2012]. Ces bases de données sont classées dans trois niveaux en fonction de leur degré de curation.

Le premier niveau contient des bases revues manuellement. Il s'agit des bases concernant *Homo sapiens*, *Escherichia coli K12*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae* et *Leishmania major*. À ces bases mono-espèce s'ajoute la seule base multi-espèces

de BioCyc : MetaCyc. Cette base de données du premier tiers de BioCyc contient l'information de 2042 voies métaboliques pour 2414 organismes et sert de base à l'inférence automatique pour les deux autres tiers de BioCyc.

Le deuxième niveau concerne des espèces pour lesquelles les données ont été obtenues par inférence électronique et qui ont subi un processus de revue manuelle moins poussé que dans le premier tiers. Parmi les 35 espèces de ce deuxième tiers, toutes sont des bactéries ou des virus, à l'exception de *Mus musculus*, *Bos taurus* et *Drosophilia melanogaster*.

Enfin le dernier niveau de BioCyc concerne les voies métaboliques de 2948 espèces de bactéries et de virus. Les données de ce dernier tiers sont issues d'inférences électroniques générées par un programme nommé PathoLogic capable de prédire les voies métaboliques d'un organisme à partir de son génome [Paley et Karp, 2002].

Les seuls vertébrés présents dans BioCyc sont donc l'Homme (niveau 1), la Souris (niveau 2) et la Vache (niveau 2). Le contenu de BioCyc est disponible en contractant une license qui est gratuite pour des besoins de recherche académique. Les données sont au format BioPAX. La faible représentation de vertébrés dans BioCyc, et notamment l'absence de la Poule, a conduit à envisager de n'utiliser BioCyc que dans le cadre d'une généralisation ultérieure à la thèse des méthodes développées à d'autres espèces.

1.1.3 KEGG

KEGG est une base de données de voies métaboliques, revues manuellement, qui concerne plusieurs espèces et qui a été développée pour l'analyse des fonctionnalités des cellules, des organismes et des écosystèmes [Kanehisa et Goto, 2000]. Elle se base sur l'information moléculaire issue de technologies expérimentales à haut-débit telles que le séquençage de génomes. KEGG répertorie 2793 espèces, dont 192 eukaryotes. Parmi ceux-ci, on compte 26 vertébrés dont l'Humain, la Souris et la Poule.

Depuis 2011, le téléchargement des données de KEGG demande de souscrire une licence payante. Ces données sont dans un format propre développé par KEGG, le format KGML. Ces deux derniers points nous ont très rapidement incité à abandonner l'utilisation de KEGG.

1.1.4 WIKIPATHWAY

Wikipathway est un projet collaboratif visant à élaborer une base de données de voies métabolique multi-espèces [Pico et al., 2008]. Wikipathway reprend d'une part les schémas de voies métaboliques disponibles dans d'autres bases de données telles que Reactome ou KEGG, et d'autre part propose des schémas créés par les utilisateurs à l'aide d'un outil d'édition graphique. Les données sont librement téléchargeables sous différents formats, dont BioPAX. En raison de sa nature collaborative, Wikipathway a une composition plus hétérogène (dans les représentations et formalismes adoptés) que les autres bases de données disponibles. Par conséquent, Wikipathway n'a été utilisé dans cette thèse qu'à des fins de recherche d'exemples et de vérifications croisées ponctuelles.

1.1.5 Ingenuity

Ingenuity Pathway Analysis (IPA) est un outil développé par Ingenuity Systems pour l'étude des voies métaboliques et réseaux biologiques [2]. Il fonctionne selon un modèle non libre payant. L'export de données générées par IPA est très limité et ne se prête pas à leur inclusion dans une étude à grande échelle. L'intérêt d'IPA dans le cadre d'une telle étude réside en la possibilité de confirmer manuellement une hypothèse particulière obtenue avec un autre outil.

1.2 Bases de connaissances et ontologies

En complément des bases de données, il existe des bases de connaissances et ontologies qui répertorient et structurent les informations relatives aux domaines qui nous intéressent. Elles constituent une ressource essentielle pour l'annotation des connaissances. Elles permettent l'application de raisonnements afin de faire apparaître des connaissances implicites à partir de celles disponibles dans les grandes bases de données.

1.2.1 Définition et propriétés d'une ontologie

Une ontologie est une représentation formelle des connaissances symboliques dans laquelle les concepts (classes) sont décrits à la fois par leur signification et par leurs relations [Bard et Rhee, 2004]. Une ontologie se présente sous la forme d'un graphe dans lequel chaque nœud est une classe relative au domaine décrit par l'ontologie. Ces nœuds peuvent être reliés par différents liens, le lien le plus fréquent étant la relation "Is a", qui relie une classe à une super-classe.

Le graphe d'une ontologie est orienté, c'est-à-dire que les relations entre les nœud ont un sens. Cela permet la description de la connaissance formalisée en allant des concepts les plus généraux aux plus précis. Dans une ontologie, une « classe » (ou « concept », ou « terme ») est un nœud du graphe. Les termes situés en amont d'un nœud sont ses « ancêtres » et ceux situés en aval sont ses « descendants ». Parmi les ancêtres d'un terme, ceux qui ne sont séparés de ce terme que par une relation sont ses « parents ». De même, parmi les descendants d'un terme, ceux qui ne sont séparés de ce terme que par une relation sont ses « enfants ». Le concept le plus général d'une ontologie n'a pas de parent ; il s'agit de la « racine ».

La figure 2 présente une ontologie très simple et non exhaustive des vertébrés. Il s'agit d'une portion de la *NCBI Taxonomy of species* [3] simplifiée pour la clareté de l'explication. Dans cet exemple, tous les termes sont liés par une relation "is a". Chaque terme a un ou plusieurs enfants et un seul parent (sauf la racine). Cette structure est celle d'un arbre, et notre ontologie est une simple taxonomie.

2. Ingenuity® Systems, `www.ingenuity.com`
3. `http://www.ncbi.nlm.nih.gov/taxonomy`

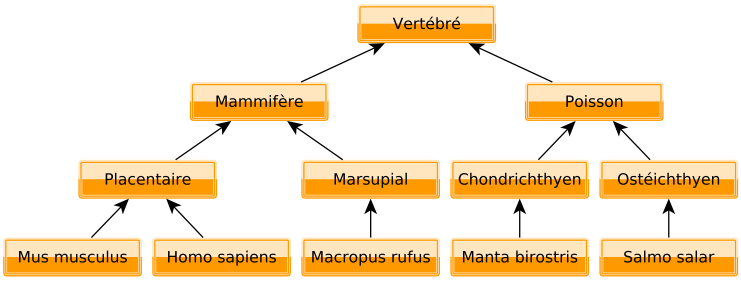


FIGURE 2 – Ontologie non exhaustive des vertébrés. Les relations sont toutes des liens “is a”. Les « vertébrés » constituent un sous-embranchement du règne animal. Il se divise en plusieurs classes, dont deux sont figurées ici : les « mammifères » et les « poissons ». Chaque classe peut être subdivisée en plusieurs groupes qui comprennent chacune des espèces.

Les concepts qui constituent les nœuds d'une ontologie peuvent être utilisés pour décrire des données par un processus d'annotation. L'intérêt d'une ontologie réside en trois propriétés importantes :

- Une ontologie est **générique**, c'est-à-dire que la connaissance qui y est formalisée est vraie tout le temps, par opposition aux données annotées, qui sont anecdoctiques. Ainsi, « Wallace est un chien » est une annotation anecdotique, alors que « les chiens sont des mammifères » est une connaissance universelle.
- Une ontologie permet le **partage** et la **réutilisation** des connaissances. En effet, une même ontologie peut servir à annoter différents jeux de données. Ainsi, la taxonomie des espèces [4] basée sur celle de Carl von Linné sert de référence à des travaux de nombreux domaines. Les principales ontologies biomédicales sont disponibles sur bioportal [Whetzel et al., 2011] ou obofoundry [5].
- Il est possible de procéder à du **raisonnement** sur une ontologie [Eiter et al., 2006]. Plusieurs types de raisonnements peuvent être appliqués, voire combinés comme la généralisation ou l'abstraction, la classification, la mesure de distance ou de similarité entre concepts ou ensembles de concepts [Jun et al., 2002; Shahar et al., 1999; Zhao et al., 2009; Wolstencroft et al., 2006; Kulik et al., 2005].

Une ontologie permet une meilleure exploitation des données stockées dans les bases de données. Cela recouvre deux types d'amélioration, qui ne sont pas exclusives. Une ontologie permet d'enrichir les requêtes afin de réduire le bruit et le silence. Une ontologie permet aussi d'interpréter les résultats d'une requête afin d'en tirer des connaissances implicites au premier abord.

Dans une ontologie, certaines relations, telle la relation “is a”, sont transitives, permettant l'héritage des ancêtres. Cela signifie que si un terme C est relié à un terme B par une relation “is a” et que B est également relié relié à A par un “is a”, alors on pourra dire que C `is a` A. Cette règle est vrai quelque soit le nombre de termes « intermédiaires ». Ainsi, dans l'ontologie donnée en exemple, *Homo sapiens* et *Mus musculus* sont tous deux des placentaires mais également des mammifères. *Macropus rufus* (le kangourou roux) est aussi un mammifère, mais par contre il n'est pas placentaire mais marsupial.

4. `http://www.ncbi.nlm.nih.gov/taxonomy/`
5. `http://www.obofoundry.org/`

En plus de la relation “is a” qui définit une hiérarchie de classes, une ontologie peut comporter des propriétés affectées à certaines classes. Dans la Figure 3, des propriétés sont associées à certaines classes. Par exemple, on peut affecter la propriété ”a la capacité de nager” à la classe « poisson ». Cette propriété s’applique alors à toutes les instances de la classes « poisson », qu’elles soient directes ou indirectes, c’est-à-dire instances d’une sous-classe de « poisson ». Puisque *Salmo salar* est une sous-classe de « poisson », on en déduit que les saumons ont la capacité de nager. Il faut remarquer qu’il s’agit ici d’une condition nécessaire (tous les poissons ont nécessairement la capacité de nager) mais pas suffisante (des animaux qui ne sont pas des poissons peuvent aussi avoir cette capacité).

Il est également possible d’affecter une propriété nécessaire et suffisante à une classe, qui agit alors comme une définition. Par exemple, on peut définir la classe « mammifère » comme l’ensemble des animaux possédant des glandes mammaires et allaitant leurs petits. Puisqu’il s’agit d’une condition nécessaire, cette définition s’applique naturellement à toutes les instances de mammifère. Le fait que ce soit également une condition suffisante permet de déduire que si un animal possède des glandes mammaires et allaite ses petits, alors c’est une instance de mammifère. Si on avait (de façon erronée) fait de la capacité de nager une définition de la classe poisson, on aurait pu en déduire que les dauphins sont des poissons. A l’inverse, la respiration exclusivement branchiale est propre aux poissons, faisant de cette propriété une condition nécessaire et suffisante (le terme « exclusivement » ayant son importance pour ne pas classer les amphibiens parmi les poissons en raison des branchies qu’ils ne possèdent qu’au stade larvaire).

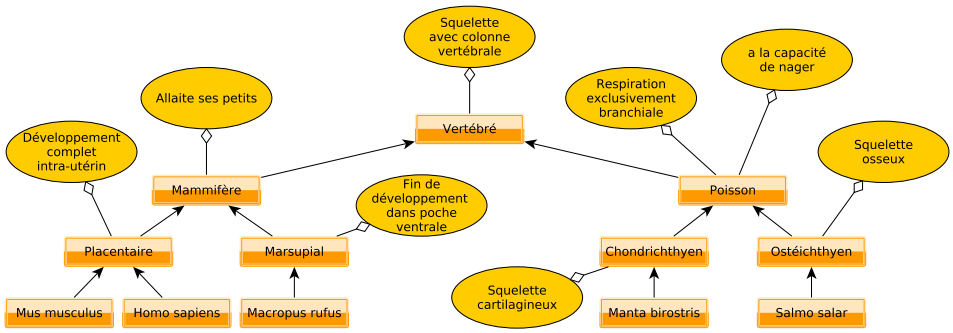


FIGURE 3 – Ontologie non exhaustive des animaux. Chaque classe peut avoir plusieurs propriétés. Ici, 7 classes sont décrites chacune par une propriété.

Il est important d’être exhaustif dans la définition des classes afin de ne pas faire d’erreur. Ainsi, si on ajoute une classe « Oiseau » à notre exemple, simplement décrite par les propriétés « possède un bec » et « est ovipare », il sera possible de classer *Ornithorhynchus anatinus* (l’ornithorynque) à la fois dans les mammifères (parce qu’il allaite ses petits) et dans les oiseaux (parce qu’il a un bec et pond des œufs). Pour éviter ce genre d’erreurs, il est possible d’utiliser la disjonction. Ainsi, dans la taxonomie des vertébrés, toutes les classes sont disjointes : il est impossible d’appartenir à plusieurs classes à la fois. Ajouter suffisamment de propriétés dans la description des classes et utiliser la disjonction à bon escient permet d’éviter les erreur.

Toutes les classes d'une ontologie ne sont pas réparties de façon homogène. On parle de différences de granularité. La figure 4 ajoute la classe « Oiseau » à notre exemple d'ontologie des vertébrés. Or cette classe n'est pas subdvisée en groupes. Les espèces qui respectent les propriétés de la classe « Oiseau » y sont directement rattachées. Seuls deux liens séparent ainsi « *Gallus gallus* » de la racine de l'ontologie, contre trois pour *Homo sapiens* : il y a une différence de granularité.

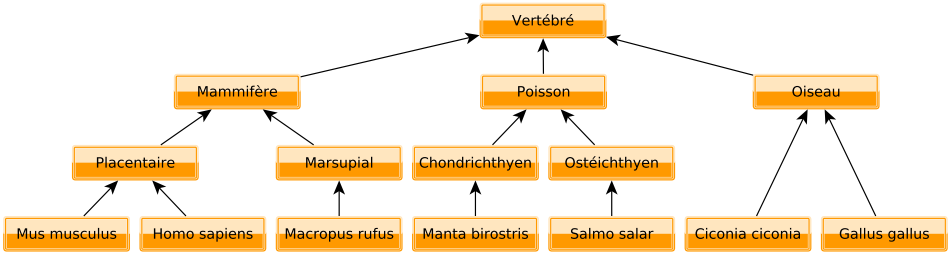


FIGURE 4 – Ontologie non exhaustive des animaux. On constate une différence de granularité entre les espèces *Ciconia ciconia* et *Gallus gallus* qui sont directement attachés à la classe taxomique des oiseaux et les autres espèces qui dépendent d'abord d'un groupe taxonomique avant d'être attaché à une classe taxonomique.

Enfin, une propriété importante des ontologies est présentée dans la Figure 5 : l'héritage multiple. A partir du moment où deux classes ne sont pas disjointes, plusieurs sous-classes peuvent s'y rattacher. Dans cette ontologie qui classifie les animaux en fonction de leur cadre de vie, on peut voir que certains animaux peuvent se trouver dans plusieurs cadres de vie différents. Ainsi, *Oryctolagus cuniculus* (le lapin) peut vivre à l'état sauvage comme être domestiqué ou élevé pour sa viande ou dans un laboratoire. Dans cet exemple, les cadres de vie ne sont pas disjoints, alors que les espèces qui y vivent le sont.

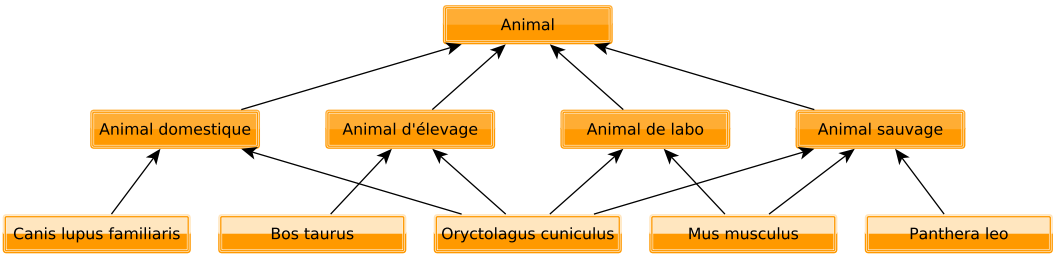


FIGURE 5 – Ontologie d'animaux classés en fonction de leur cadre de vie. Chaque espèce peut se trouver dans différent cadres de vie.

1.2.2 GENE ONTOLOGY

Gene Ontology (GO) est un projet visant à standardiser la représentation des connaissances concernant les gènes et produits de gènes [Ashburner et al., 2000]. GO propose un vocabulaire contrôlé, composé de termes hierarchisés et permettant de décrire les ca-

ractéristiques d'un produit de gène. Ce vocabulaire est commun à tous les produits de gènes, quels que soient les gènes et les espèces considérés. GO est divisé en trois sections principales indépendantes relatives aux processus biologiques (*biological process*, BP), aux fonctions moléculaires (*molecular functions*, MF) et aux composants cellulaires (*cellular component*, CC).

Les nœuds de Gene Ontology sont des termes décrivant les caractéristiques d'un produit de gène. Ils sont appelés "Termes GO". Ces termes GO sont liés par cinq relations différentes :

- "Is a" est une relation simple de type classe/sous-classe. A `is a` B signifie que A est une sous-classe de B, c'est-à-dire que toutes les instances de A sont des instances de B. Si A `is a` B `is a` C, on peut inférer que A `is a` C.
- "Part of" est une relation de composition partielle. C `part of` D signifie que chaque instance de C est toujours une partie d'au moins une instance de D. Cela n'implique pas que toutes les instances de D aient au moins une partie qui soit une instance de C. Si A `part of` B `part of` C, alors A `part of` C.
- La relation "Regulates" et ses 2 sous-relations "Positively Regulates" et "Negatively Regulates" décrivent une interaction entre un processus biologique et un autre. A `Regulates` B signifie que chaque instance de A régule B, mais que toutes les instances de B ne sont pas forcément régulées par A. Si A `regulates` B `is a` C, ou bien si A `is a` B `regulates` C, alors A `regulates` C. Il en va de même pour les relations Positively et Negatively Regulates.

La figure 6 présente un extrait de GO.

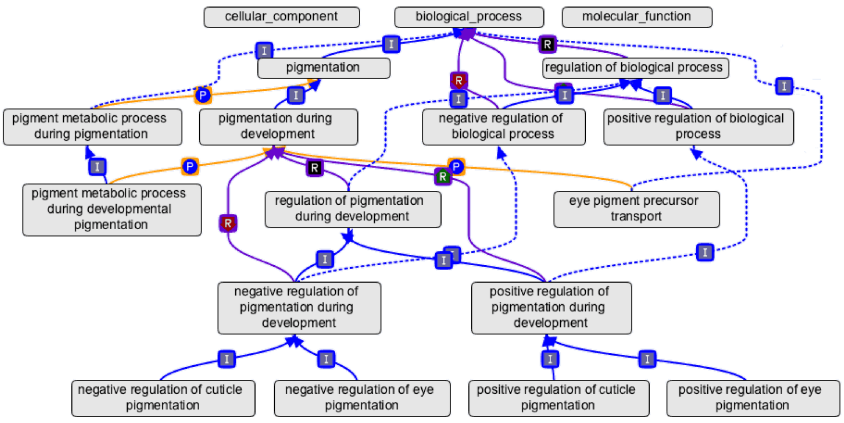


FIGURE 6 – Extrait de Gene Ontology. Les relations entre les termes sont représentées par les flèches colorées. L'initiale du nom de la relation figure sur la flèche (I : `is a`, P : `part of`, R sur fond noir : `regulates`, R sur fond rouge : `negatively regulates` et R sur fond vert : `positively regulates`). Cette image est issue de la documentation du site web de GO.

1.2.3 GENE ONTOLOGY ANNOTATION

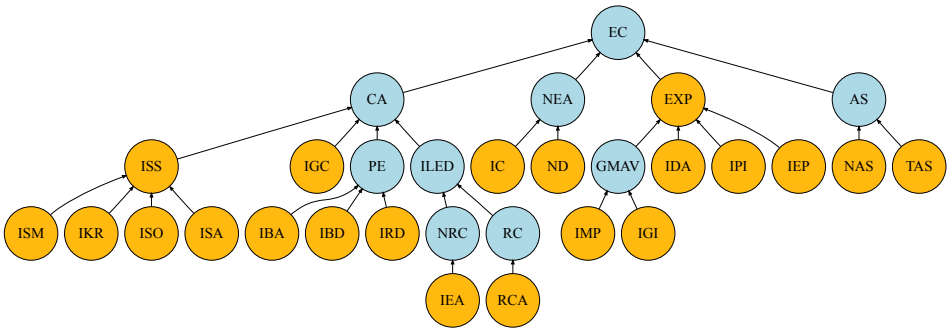
Gene Ontology Annotation (GOA) est un projet du *European Bioinformatics Institute* (EBI) ayant pour but l'annotation de produits de gènes de différentes espèces par des

termes GO [Camon et al., 2003]. Il se base sur plusieurs bases de données comme UniProt ou Ensembl, chaque entrée restant unique. GOA est donc un trait d'union entre ces bases de données et Gene Ontology [Hill et al., 2008]. Chaque produit de gène est identifié dans GOA par son symbole et son numéro de taxon, ainsi que par un id propre à chaque base de données de gènes. C'est par le biais de cette identification que chaque produit de gène est associé à un ou plusieurs termes GO.

La base de données GOA propose des tables séparées pour les annotations de produits de gènes de 7 espèces modèles (Humain, Souris, Rat, Arabidopsis, Poule, Vache et Poisson Zèbre) ainsi que celles de produits de gènes répertoriés dans diverses bases de données inter-espèces (PDB, UniProt, Proteomes...).

La façon dont un terme GO a été associé à un produit de gène au cours du processus d'annotation est précisée par un "Evidence Code" (EC). Il en existe actuellement 21 différents. Ces EC sont séparés en 5 catégories principales de niveau de preuve : expérimental (*Experimental EC*), computationnel (*Computational Analysis EC*), déclaration d'auteur (*Author Statement EC*), déclaration de correcteur (*Curator Statement EC*) et annotation automatique (*Automatically-assigned EC*). Tous ces niveaux sont subdivisés en EC plus précis sauf le dernier qui ne contient que le code Inferred from Electronic Annotation (IEA), qui est le seul code qui qualifie une annotation non vérifiée par un correcteur. La Figure 7 présente les evidence codes de GO organisés dans une ontologie. Nous avons ajouté des catégories intermédiaires (en bleu) pour construire cette ontologie que nous avons utilisée par la suite.

Gene Ontology précise que les Evidence Codes ne sont pas des indicateurs de la qualité des annotations, et ne doivent par conséquent pas être utilisés comme une mesure de cette qualité. Cependant, il est aussi précisé que dans chaque catégorie de codes, les méthodes utilisées produisent des annotations de plus ou moins haut niveau de confiance et spécificité. Il résulte de ce point de vue que les annotations associées à un EC expérimental sont généralement considérées comme étant de meilleure fiabilité que les autres, bien que cela n'ait pas été démontré [Rhee et al., 2008]. Il faut de plus souligner que l'annotation automatique (code IEA) représente 93.67% de la totalité des annotations présentes dans la table multi-espèces de GOA basée sur les identifiants UniProt. Ce taux d'annotations inférrées automatiquement varie entre les espèces. Les Figures 8, 9 et 10 montrent la répartition des evidence codes dans l'annotation respective de la poule, de la souris et de l'humain.



Catégories présentes dans Gene Ontology	
Experimental Evidence Codes	EXP: Inferred from Experiment
	IDA: Inferred from Direct Assay
	IPI: Inferred from Physical Interaction
	IMP: Inferred from Mutant Phenotype
	IGI: Inferred from Genetic Interaction
	IEP: Inferred from Expression Pattern
Computational Analysis Evidence Codes	ISS: Inferred from Sequence or Structural Similarity
	ISO: Inferred from Sequence Orthology
	ISA: Inferred from Sequence Alignment
	ISM: Inferred from Sequence Model
	IGC: Inferred from Genomic Context
	IBA: Inferred from Biological aspect of Ancestor
	IBD: Inferred from Biological aspect of Descendant
	IKR: Inferred from Key Residues
	IRD: Inferred from Rapid Divergence
	RCA: inferred from Reviewed Computational Analysis
Author Statement Evidence Codes	TAS: Traceable Author Statement
	NAS: Non-traceable Author Statement
Curator Statement Evidence Codes	IC: Inferred by Curator
	ND: No biological Data available
Automatically-assigned Evidence Codes	IEA: Inferred from Electronic Annotation

Catégories ajoutées pour l'ontologie des evidence codes
EC: Evidence Code
GMAV: Genetic Mutation or Allelic Variation
NEA: No Evidence Available
AS: Author Statement
CA: Computational Analysis
PE: Phylogenetic Evidence
ILED: Inferred from Large-scale Experimental Dataset
NRC: Not Reviewed by Curator
RC: Reviewed by Curator

FIGURE 7 – Ontologie des evidence codes de GO. Les nœuds des EC utilisés dans GO sont jaunes, les nœuds ajoutés pour l'organisation de cette ontologie sont bleus.

L'association d'un terme à un gène peut être précisée par un "qualifier".

- Le qualifier **NOT** dit que le gène **n'est pas associé** avec le terme GO mentionné. Dans le cas où une protéine P présente par exemple une forte similarité de séquence avec une enzyme E, mais qu'il a été prouvé expérimentalement que P n'a aucune activité enzymatique, le fait de l'annoter avec le terme ≪ activité enzymatique ≫ et le qualifier **NOT** empêche que l'annotation automatique associe ≪ activité enzymatique ≫ à un autre produit de gène présentant également une forte similarité de séquence avec l'enzyme E et la protéine P. Par conséquent, aucun terme associé au code IEA n'est précisé par le qualifier NOT, les potentiels faux positifs étant éliminés en amont de l'annotation par inférence électronique.
- "Colocalizes with" ne concerne que les termes de la branche Cellular Component, et signifie que le produit du gène est associé à un organite ou à un complexe.
- "Contributes to" ne concerne que les termes de la branche Molecular Function, et signifie que le produit du gène prend part à un complexe. Ce complexe appartenant à un cadre cellulaire, un gène ayant un terme qualifié par "Contributes to" doit avoir également un terme classé dans Cellular Component décrivant le complexe.

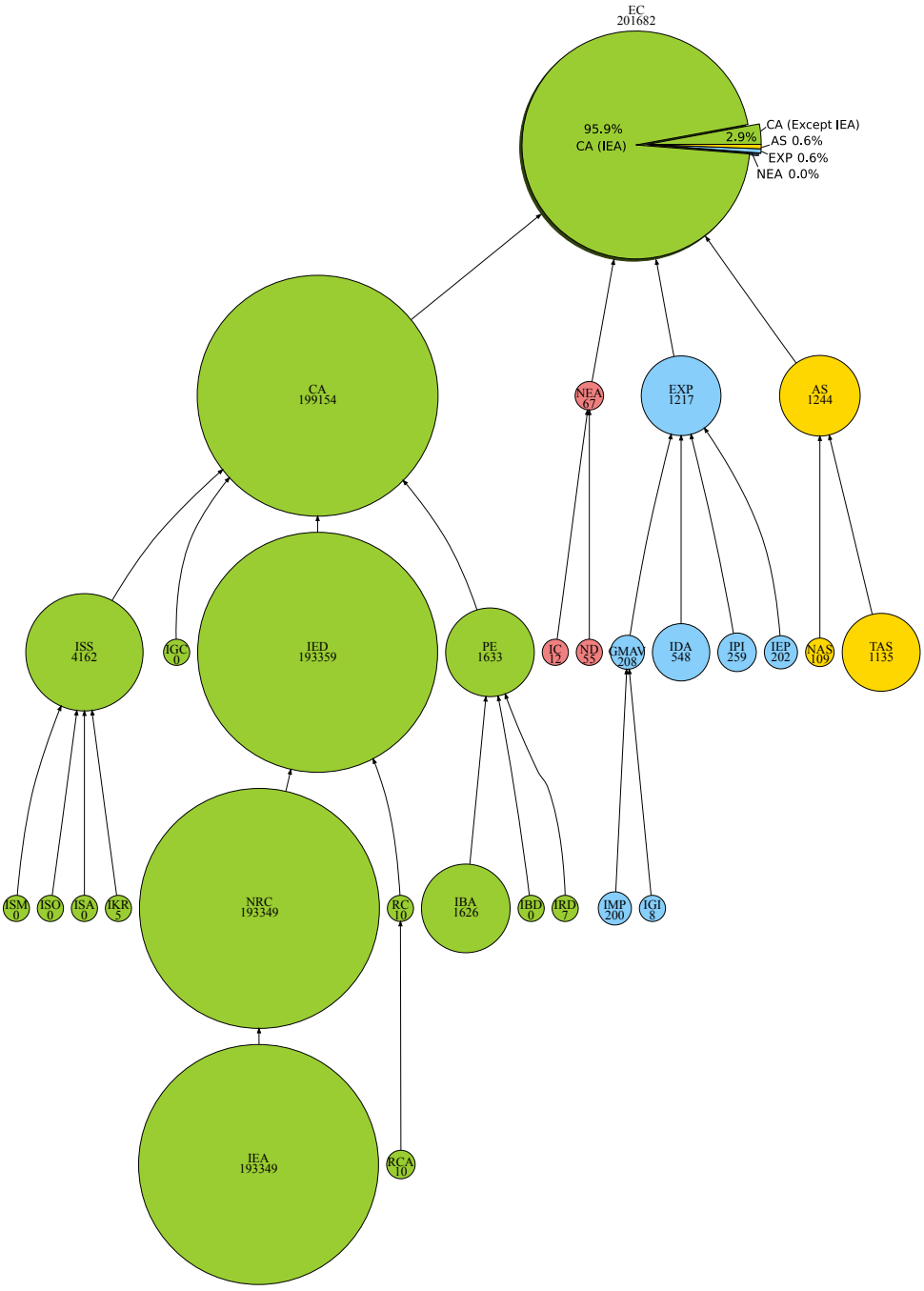


FIGURE 8 – Répartition des EC chez la poule. 95.9% de l'annotation de la poule provient d'un processus d'inférence automatique non revue manuellement par un curateur.

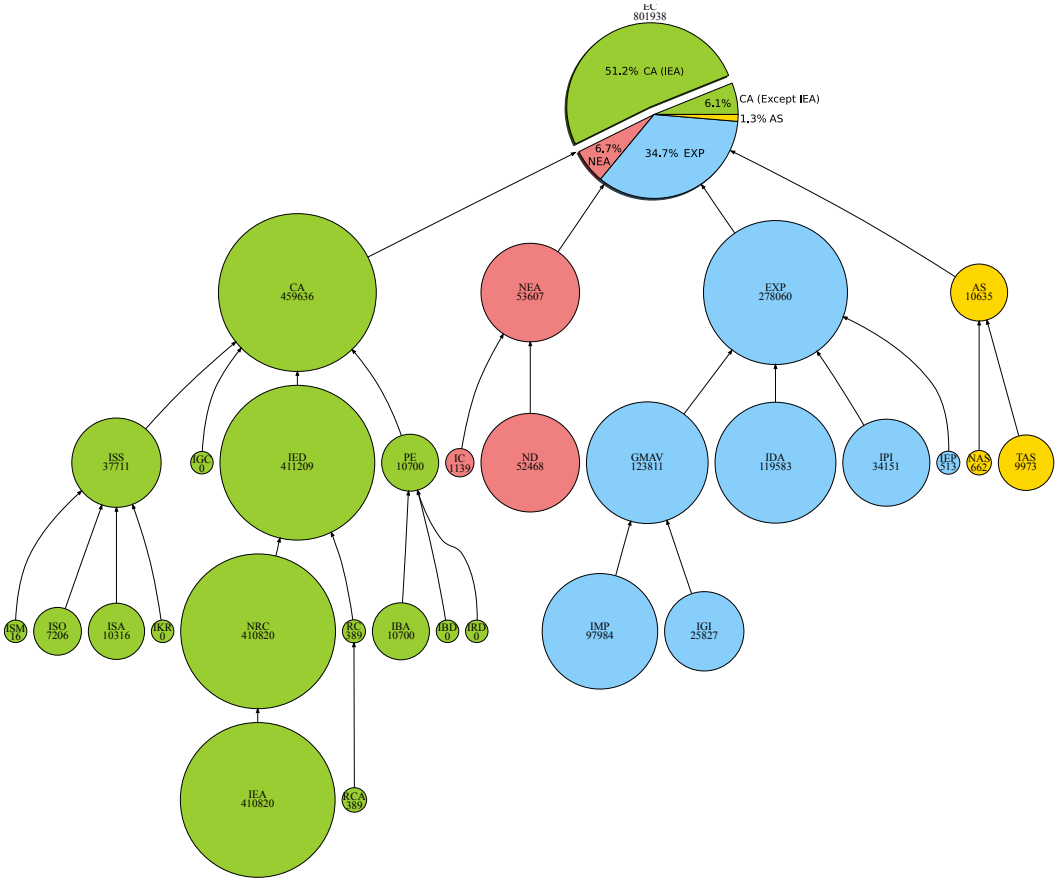


FIGURE 9 – Répartition des EC chez la souris. 51.2% de l'annotation de la souris provient d'un processus d'inférence automatique non revue manuellement par un curateur.

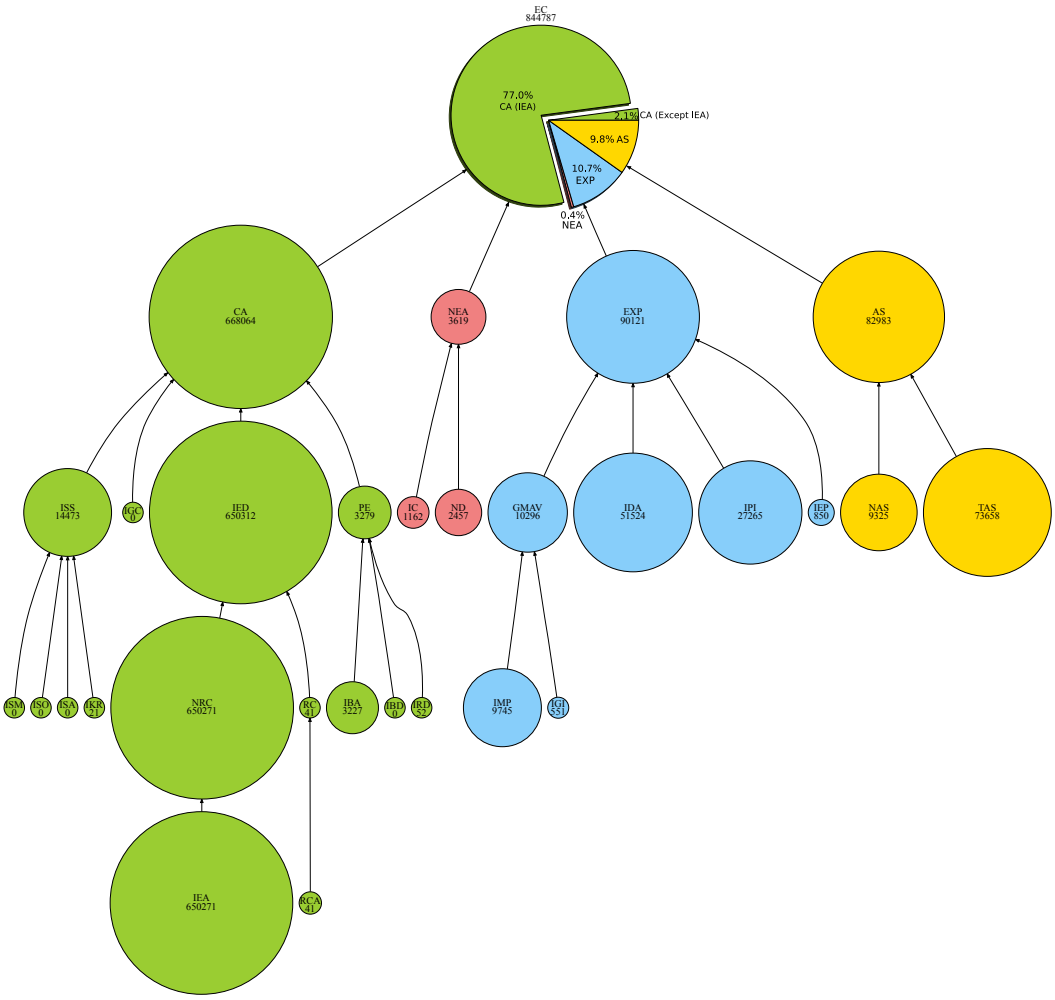


FIGURE 10 – Répartition des EC chez l'humain. 77% de l'annotation de l'humain provient d'un processus d'inférence automatique non revue manuellement par un curateur.

2 Méthodes de comparaison de termes et d'ensembles de termes d'une ontologie

Notre objectif était de développer une méthode pour comparer fonctionnellement les voies métaboliques entre espèces sur la base des annotations des produits de gènes qui y interviennent. Ces annotations sont disponibles dans Gene Ontology Annotation. Chaque gène peut être annoté par plusieurs termes de Gene Ontology. Il nous fallait donc utiliser une approche permettant de comparer des ensembles de termes d'une ontologie afin de quantifier la similarité entre ces ensembles.

2.1 Métriques simples : Jaccard et Dice

L'index de Jaccard est le rapport entre la taille de l'intersection des ensembles considérés et la taille de l'union des ensembles. L'équation 1 permet de calculer l'index de Jaccard des ensembles A et B.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Le cœfficient de Dice est le rapport entre le double de la taille de l'intersection des ensembles considérés et la taille de l'union des ensembles. L'équation 2 permet de calculer le cœfficient de Dice des ensembles A et B.

$$D(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{2}$$

Il est possible de convertir ces deux métriques à l'aide de la formule 3 :

$$D = \frac{2 \times J}{1 + J} \tag{3}$$

Ces deux métriques ensemblistes sont bien adaptées pour calculer des similarités entre des éléments indépendants les uns des autres et équiprobables. Ce n'est pas le cas des annotations GO, qui ne vérifient aucun de ces deux principes. En effet, les termes GO ne sont pas indépendants puisque chaque terme hérite de l'information contenue dans ses ancêtres. Ils ne sont pas non plus équiprobables, parce que certains termes GO annotent plus de produits de gènes que d'autres termes GO de même précision [Mazandu et Mulder, 2012].

2.2 Mesures de distances et similarités sémantiques

Afin de prendre en compte les propriétés d'une ontologie, des mesures plus complexes ont été développées pour comparer des termes et des ensembles de termes. On parle de mesures de distances et de similarités sémantiques. Pesquita et al. [2009] ont procédé à une revue de ces mesures, qui se déclinent en trois catégories selon qu'elles sont basées sur un comptage d'arêtes, sur une valeur attribuée aux nœuds, ou sur une combinaison des deux. La plupart des mesures présentées ci-après ne concernent que la similarité

entre deux termes, et non entre deux ensembles de termes. Or lorsqu'on souhaite obtenir la similarité sémantique entre deux gènes, on a besoin de comparer les deux ensembles X et Y constitués par les termes qui les annotent. Lors d'une comparaison d'ensembles de termes, il faut donc calculer la similarité de chaque terme du premier ensemble avec chaque terme du deuxième. La similarité des ensembles peut ensuite être obtenue de trois façons :

- En calculant la moyenne des résultats de toutes ces comparaisons entre termes [Lord et al., 2003].
- En calculant cette même moyenne en ne considérant pour chaque terme que sa plus haute valeur de similarité lorsqu'on le compare à l'autre ensemble [Couto et al., 2007; Azuaje et al., 2006; Wang et al., 2007]. Un exemple de ce mode de calcul est donné dans les équations 13 et 14 plus loin dans ce document.
- En prenant le maximum de tous les résultats des comparaisons entre termes [Sevilla et al., 2005].

2.2.1 Méthodes basées sur les arêtes

1. **Distance de Rada**
   La distance de Rada entre deux termes A et B appartenant à un graphe est égale au nombre minimal d'arrêtes qui les sépare [Rada et al., 1989]. Cette mesure est comparable à l'algorithme de Dijkstra qui calcule le plus court chemin entre deux points d'un graphe [Dijkstra, 1959].

2. **Mesure de Wu et Palmer**
   La mesure de Wu et Palmer [1994] compare deux termes A et B au sein d'un graphe de racine R en utilisant trois distances : D1 la distance entre A et R, D2 la distance entre B et R et D la distance entre le plus proche ancêtre commun de A et de B et R. La figure 11 illustre l'équation 4.

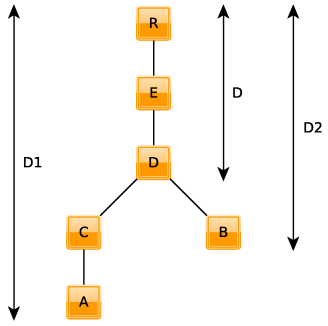


FIGURE 11 – Illustration de la mesure de Wu et Palmer.

$$Sim(A,B) = \frac{2 * D}{D1 + D2} \tag{4}$$

3. **Mesure de Pekar et Staab**
   Cette mesure utilise des distances D1 et D2 correspondant au plus long chemin entre les termes à comparer A et B et leur plus proche ancêtre commun C, ainsi qu'une distance D correspondant au plus long chemin entre C et la racine R [Pekar et Staab, 2002]. Le calcul est présenté dans l'équation 5. Cette mesure a été utilisée sur GO par Yu et al. [2005].

$$Sim(A, B) = \frac{D}{D + D1 + D2} \tag{5}$$

2.2.2 Méthodes basées sur les nœuds

1. **Mesure de Resnik**
   Resnik a développé sa mesure pour calculer la similarité entre des termes de WordNet [Resnik, 1999]. Il attribue à chaque terme une valeur qui représente la quantité d'information qu'il contient. On parle de "Contenu d'information" (*Information content*, IC). Ce concept est utilisé par toutes les autres mesures basées sur les nœuds. L'IC d'un terme dépend de la probabilité de le rencontrer. Plus un terme est rare, plus il est considéré comme informatif. Et inversement, plus la probabilité de rencontrer un terme est faible, plus son IC est élevé. L'IC du terme A est calculé en appliquant l'équation 6 :

$$IC(A) = -\log P(A) \tag{6}$$

   La similarité sémantique entre deux termes A et B est l'IC de leur ancêtre commun le plus informatif (*Most Informative Common Ancestor*, MICA).

2. **Mesure de Jiang/Conrath**
   Cette mesure se base également sur l'IC, mais à la différence de Resnik, elle prend en compte la distance entre les termes comparés et leur MICA [Jiang et Conrath, 1997].

$$Sim_{JC}(A, B) = 1 - IC(A) + IC(B) - 2 \times IC(MICA) \tag{7}$$

3. **Mesure de Lin**
   Comme la précédente, la mesure de Lin utilise l'IC en tenant compte de la distance entre les termes comparés et leur MICA [Lin, 1998].

$$Sim_{Lin}(A, B) = \frac{2 \times IC(MICA)}{IC(A) + IC(B)} \tag{8}$$

4. **Mesure de Schliker**
   Cette mesure pondère la mesure de Lin par la probabilité du MICA afin de faire en sorte que le résultat varie selon la position des termes A et B dans le graphe et non

seulement selon la distance entre ces termes et leur ancêtre commun [Schlicker et al., 2006].

$$Sim_{Sch}(A,B) = Sim_{Lin}(A,B) \times (1 - log(P(MICA))) \quad (9)$$

5. **Mesure de Couto**
   Pour prendre en compte le fait que deux termes peuvent avoir plusieurs ancêtres communs disjoints (*Disjoint Common Ancestors*, DCA), Couto applique toutes les mesures précédemment proposées en remplaçant l'IC du MICA par la moyenne de l'IC de tous les DCA [Couto et al., 2007].

2.2.3 Méthodes hybrides

Il est possible de combiner les deux principes précédents dans une mesure hybride.

1. **Mesure de Wang**
   Wang attribue à chaque terme une valeur sémantique [Wang et al., 2007]. Cette démarche relève de la catégorie des mesures basées sur les nœuds. Cependant, cette valeur sémantique est elle-même calculée en parcourant le graphe, ce qui relève de la catégorie des mesures basées sur les arêtes.

   Prenons un exemple concret pour expliquer le fonctionnement de la méthode de Wang. Soient deux ensembles de termes GO "Set 1" et "Set 2" qui annotent deux gènes différents que l'ont souhaite comparer (Figure 12). Ils n'ont chacun que 33% d'annotations communes. La figure 13 place ces termes dans le graphe de GO et la Figure 14 procède à l'extention aux ancêtres de chaque terme. Chaque terme hérite en effet de la connaissance portée par ses ancêtres. La proportion d'annotations communes devient 62.5% pour le Set 1 et 83.3% pour le Set 2.

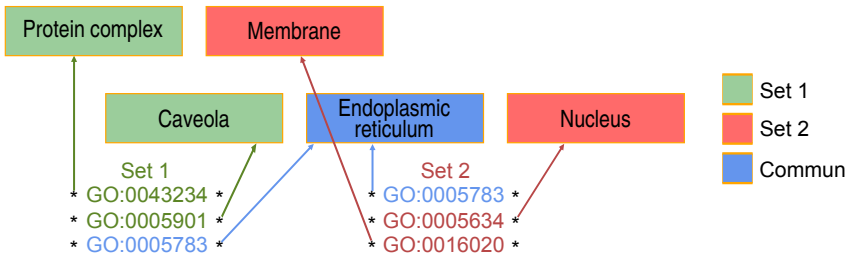


FIGURE 12 – Set 1 et Set 2 sont deux ensembles de termes GO annotant deux gènes.

Wang commence par calculer les contributions sémantiques des ancêtres de chacun des termes à comparer (équation 10).

$$\begin{cases} S_A(A) = 1 \\ S_A(t) = max\{w_e * S_A(t') \mid t' \in children\ of\ (t)\}\ if\ t \neq A \end{cases} \quad (10)$$

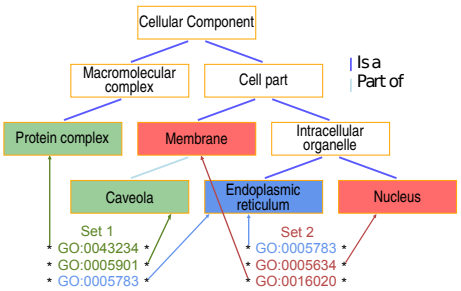


FIGURE 13 – Les termes de Set 1 et Set 2 sont hierarchisés dans GO.

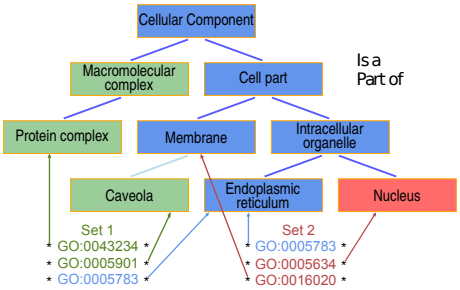


FIGURE 14 – Les ancêtres des termes de Set 1 et Set 2 appartiennent implicitement à ces ensembles.

Dans cette équation, $S_A(t)$ est la contribution sémantique du terme t au terme A et w_e est le facteur de contribution sémantique pour l'arrête e qui relie un terme t à son enfant t'. Wang a défini les facteurs de contribution sémantiques suivants : w_{isa} = 0,8 et w_{partof} = 0,6. Ensuite, la valeur sémantique de chaque terme à comparer est définie d'après l'équation 11. Ce calcul est illustré pour les termes "*Caveola*" et "*Endoplasmic reticulum*" de notre exemple dans les Figures 15 et 16.

$$SV(A) = \sum_{t \in T_A} S_A(t) \tag{11}$$

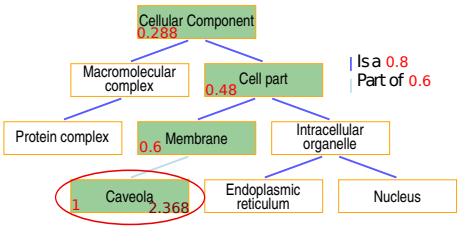


FIGURE 15 – La valeur sémantique de "*Caveola*" est égale à la somme des contributions sémantiques de ses ancêtres, soit 2.368.

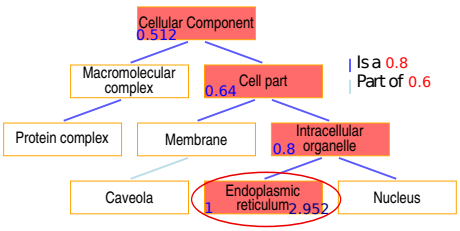


FIGURE 16 – La valeur sémantique de "*Endoplasmic reticulum*" est égale à la somme des contributions sémantiques de ses ancêtres, soit 2.952.

La comparaison entre deux termes A et B se calcule avec l'équation 12 :

$$S_{GO}(A,B) = \frac{\sum_{t \in T_A \cap T_B} (S_A(t) + S_B(t))}{SV(A) + SV(B)} \tag{12}$$

Ainsi, la similarité entre "Caveola" et "Endoplasmic reticulum" est de 0.36 (Figure 17).

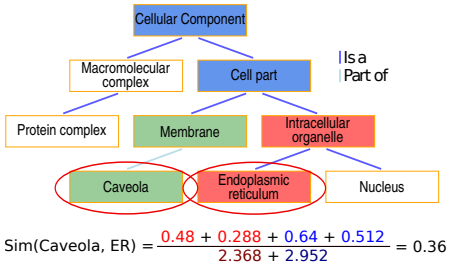


FIGURE 17 – La similarité sémantique entre "*Caveola*" et "*Endoplasmic reticulum*" est égale au rapport de la somme des contributions sémantiques de leurs ancêtres communs et de la somme de leurs valeurs sémantiques.

La similarité entre un terme "go" et un ensemble de termes "GO" se calcule avec l'équation 13 :

$$Sim(go, GO) = \max_{1 \leqslant i \leqslant k} (S_{GO}(go, go_i)) \tag{13}$$

Ainsi, la similarité entre "Caveola" et "Set 2" est de 0.78 (Figure 18).

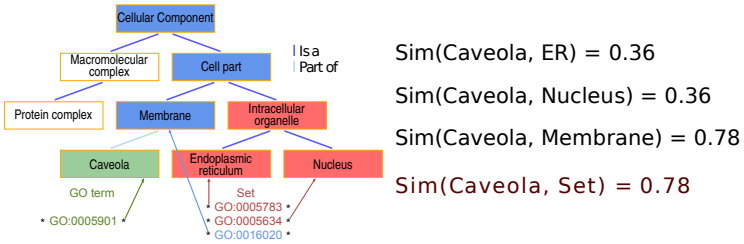


FIGURE 18 – La similarité sémantique entre "*Caveola*" et Set 2 est égale à la plus grande similarité sémantique mesurée entre "*Caveola*" et chacun des termes appartenant à Set 2.

Enfin, la similarité entre deux gènes se calcule avec l'équation 14, illustrée pour notre exemple par la Figure 19. Le résultat de la mesure de Wang étant borné entre 0 et 1, la valeur de 0.75 obtenue est bien plus forte que les 33% d'annotations communes que l'on trouvait initialement pour les deux ensembles. Comme elle tient compte de la contribution sémantique des ancêtres des termes comparés, cette valeur de similarité représente mieux la réalité.

$$Sim(G1, G2) = \frac{\sum_{1 \leqslant i \leqslant m} (Sim(go_{1i}, GO_2) + \sum_{1 \leqslant j \leqslant n} (Sim(go_{2j}, GO_1)}{m + n} \tag{14}$$

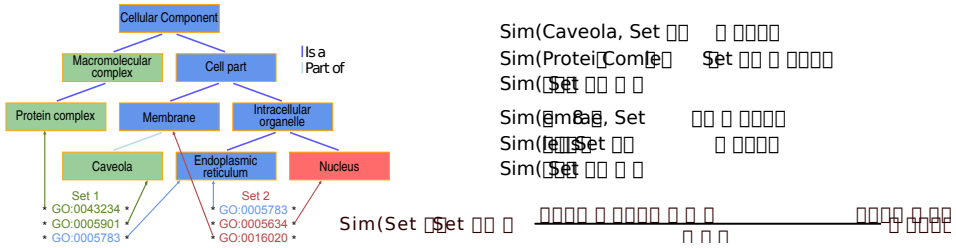


FIGURE 19 – La similarité sémantique entre Set 1 et Set 2 est égale au rapport entre la somme de la similarité de chaque terme de Set 1 avec Set 2 et de la similarité de chaque terme de Set 2 avec Set 1 et la somme du nombre de termes présents dans Set 1 et Set 2, soit 0.75.

2. **Mesure de Othman**
   Othman et al. ont proposé une mesure de distance hybride dans laquelle chaque arête est pondérée par la profondeur des nœuds, la densité des liens pour chaque nœuds comparé et la différence entre l'IC des nœuds comparés [Othman et al., 2008].

3 SYNTHÈSE

La comparaison inter-espèces de voies métaboliques repose sur une ou plusieurs bases de données contenant la succession des réactions chez les espèces à comparer. Des produits de gènes interviennent tout au long de chaque voie métabolique, la plupart en tant qu'enzyme catalysant une réaction. Ces produits de gènes sont annotés par des termes de Gene Ontology, ce qui permet de les comparer entre eux à l'aide d'une mesure de similarité sémantique. Les deux métriques simples (Dice et Jaccard) présentées dans cette section permettent de comparer deux ensembles pour en évaluer la similarité. Elles sont parfois utilisées pour comparer des termes appartenant à une ontologie, ce qui est une erreur [Rhee et al., 2008]. En effet, ces métriques ne tiennent pas compte de la notion d'héritage inhérente à une ontologie. Il faut donc sélectionner une mesure sémantique. Cette mesure doit supporter la comparaison de gènes entre espèces. Cette condition n'est pas respectée par les méthodes basées sur l'IC des termes GO. En effet, l'IC d'un terme dépend de la probabilité qu'il annote un gène. Cette probabilité est calculée par la fréquence à laquelle le terme annote un gène. Il est possible de calculer cette fréquence sur l'annotation de chaque espèce, menant à autant de valeurs d'IC pour chaque terme qu'il y a d'espèces et empêchant la comparaison inter-espèces. Il est également possible de calculer cette fréquence en cumulant toutes les annotations de toutes les espèces, mais cela conduit à un fort biais en faveur des caractéristiques les mieux connues des espèces les plus étudiées. La comparaison inter-espèce est possible avec les méthodes basées sur les arêtes, puisqu'elles ne dépendent pas d'un corpus d'annotations. Cependant, le fait que les termes GO ne soient pas distribués de manière homogène dans l'ontologie affaiblit la pertinence des résultats obtenus par ces méthodes. La méthode hybride de Wang est celle qui se rapproche le plus d'une méthode basée sur les nœuds sans être dépendante d'un corpus d'annotations.

Toutes les mesures de similarité présentées dans ce chapitre ont été évaluées [Pesquita et al., 2009; Couto et al., 2007] Il en résulte qu'elles sont globalement performantes pour déterminer la similarité entre deux gènes. Cependant, elle sont capables d'attribuer à deux gènes g1 et g2 une similarité haute très proche de celle qu'elles attribuent à deux autres gènes g3 et g4 à partir du moment où g1 et g2, comme g3 et g4 ont suffisamment d'annotations en commun, et ce même s'il s'avère que dans une de ces paires de gènes, un gène a en plus des annotations spécifiques qui traduisent des caractéristiques biologiques particulières. Cette situation de gènes similaires dont au moins un a des particularités est biologiquement intéressante. On souhaite les distinguer des gènes similaires n'ayant pas de fonctions particulières. Le problème est que cette situation est relativement rare et qu'il faut être capable de l'identifier parmi la masse des données similaires. Il existe donc un besoin de quantifier la particularité d'un gène à l'aide d'une nouvelle mesure sémantique.

Deuxième partie

Résultats

CHAPITRE 3

PARTICULARITÉ SÉMANTIQUE

Dans ce chapitre, nous présentons une nouvelle mesure de particularité sémantique développée dans le cadre de cette thèse. Cette mesure a pour but de quantifier la part des processus, fonctions et localisations qui sont spécifiques à un gène lorsqu'on le compare à un autre. Elle aurait pu s'appeler « mesure de spécificité », mais ce terme désigne aussi une mesure statistique très employée et aurait donc conduit à une certaine confusion. Cette mesure de particularité a fait l'objet d'un article soumis à PLoS ONE, présenté dans ce chapitre.

Sommaire

1 Introduction

Dans la présentation du contexte biologique, nous avons exposé les grandes lignes du métabolisme des lipides chez l'Homme et chez la poule. Ce métabolisme présente des particularités chez chacune des deux espèces. Ce sont ces partcularités qui nous intéressent. Dans l'état de l'art, nous avons conclu au chapitre 2 que l'utilisation de mesures de similarité sémantique seules ne permettait ni d'identifier ni de quantifier correctement les particularités d'un gène lorsqu'on le compare à un autre avec lequel il a de nombreuses fonctions en commun.

Pour répondre à ce besoin, nous avons donc développé une nouvelle méthode de mesure sémantique. Nous l'avons appelée « particularité sémantique ». Cette mesure permet de comparer les annotations de deux produits de gènes en se focalisant sur les termes GO spécifiques à chacun des produits de gènes comparés. Pour tenir compte des différences de granularités possibles entre annotations, elle repose sur la notion d'informativité. Notre mesure de particularité calcule le rapport de l'informativité particulière ("*particular informativeness*", PI) d'un gène d'intérêt et de l'informativité commune entre ce gène et celui auquel on souhaite le comparer. L'informativité d'un terme est une notion générique qui permet avec notre mesure de fonctionner avec plusieurs métriques différentes. Ainsi, lorsque qu'un corpus est disponible pour calculer un contenu d'information (IC), cet IC peut

être utilisé comme informativité. A l'inverse, s'il est impossible d'obtenir un tel corpus, par exemple dans le cadre d'une comparaison inter-espèces, la valeur sémantique de Wang peut être utilisée comme informativité.

Cette étude a fait l'objet de l'article intitulé "Semantic particularity measure for functional characterization of gene sets using Gene Ontology" accepté pour publication dans PLoS ONE.

2 Article

Semantic particularity measure for functional characterization of gene sets using Gene Ontology

- Résumé -

Contexte : Le traitement de données génétiques et génomiques résulte souvent en la construction d'importants ensembles de gènes. La comparaison fonctionnelle de ces ensembles de gènes est une des clés de l'analyse de ces données, via l'identification des fonctions communes à un ensemble gènes et de celles qui diffèrent. Gene Ontology fournit un vocabulaire de référence pour l'analyse des fonctions moléculaires, des processus biologiques et des composants cellulaires dans lesquels les gènes sont impliqués. De nombreuses mesures de similarité sémantique ont été développées pour quantifier systématiquement l'importance des termes GO communs à deux gènes. Cet article présente comment la comparaison d'ensembles de gènes peut être améliorée en considérant la particularité sémantique de gènes au sein d'un ensemble de gènes en complément de la similarité sémantique.

Résultats : Nous proposons une nouvelle approche pour calculer les particularités sémantiques d'un ensemble de gènes basée sur l'informativité des termes GO. L'informativité d'un terme GO peut être calculée soit à partir de son contenu d'information ("*information content*", IC) basé sur la fréquence de ce terme au sein d'un corpus, soit sur une fonction de la distance de ce terme à la racine de l'ontologie. Nous avons défini la particularité d'un ensemble de terme GO Sg1 comparé à un autre ensemble Sg2 à partir de cette informativité. Nous avons combiné notre mesure de particularité avec une mesure de similarité pour comparer des ensembles de gènes. Nous avons démontré que cette combinaison est capable d'identifier des gènes ayant des fonctions particulière au sein d'ensembles de gènes similaires. L'utilisation seule d'une mesure de similarité ne permet pas cette identification.

Conclusion : La particularité sémantique devrait être utilisée en conjonction d'une mesure de similarité sémantique afin de procéder à l'analyse fonctionnelle d'ensembles de gènes annotés par des termes GO. Le principe de la particularité sémantique est généralisable à d'autres ontologies.

Semantic particularity measure for functional characterization of gene sets using Gene Ontology

Charles Bettembourg[1,2,3,4,5*], Christian Diot[2,3], Olivier Dameron[1,4,5]
1 Université de Rennes 1, 35000 Rennes, France
2 INRA, UMR1348 PEGASE, Saint-Gilles, France
3 Agrocampus OUEST, UMR1348 PEGASE, Rennes, France
4 IRISA, Campus de Beaulieu, 35042 Rennes, France
5 INRIA, France
∗ **E-mail: charles.bettembourg@univ-rennes1.fr**

Abstract

Background: Genetic and genomic data analyses are outputting large sets of genes. Functional comparison of these gene sets is a key part of the analysis, as it identifies their shared functions, and the functions that distinguish each set. The Gene Ontology (GO) initiative provides an unified reference for analyzing the genes molecular functions, biological processes and cellular components. Numerous semantic similarity measures have been developed to systematically quantify the weight of the GO terms shared by two genes. We studied how gene set comparisons can be improved by considering gene set particularity in addition to gene set similarity.

Results: We propose a new approach to compute gene set particularities based on the information conveyed by GO terms. A GO term informativeness can be computed using either its information content based on the term frequency in a corpus, or a function of the term's distance to the root. We defined the semantic particularity of a set of GO terms Sg1 compared to another set of GO terms Sg2. We combined our particularity measure with a similarity measure to compare gene sets. We demonstrated that the combination of semantic similarity and semantic particularity measures was able to identify genes with particular functions from among similar genes. This differentiation was not recognized using only a semantic similarity measure.

Conclusion: Semantic particularity should be used in conjunction with semantic similarity to perform functional analysis of GO-annotated gene sets. The principle is generalizable to other ontologies.

Introduction

With the continued advance of high-throughput technologies, genetic and genomic data analyses are outputting large sets of genes. The amount of data involved requires automated comparison methods [1]. The characterization of these sets typically consists in a combination of the following three operations [2, 3]: first, synthesize the over- and under-represented functions of these genes [4, 5]; second, identify how these genes interact with each other [6]; third, identify and quantify the common shared features and the differentiating features [7,8]. A widely used method for genes sets study called "Gene Set Enrichment Analysis" (GSEA) determines which gene features are over-represented in a gene set [9]. Numerous tools have been developed in this purpose: BiNGO [10], GOEAST [11], ClueGO [12], DAVID [13], GeneWeaver [14], GOTM [15]. See Hung et al. recent work for a review [16]. GSEA is useful for clustering a set of genes into subsets sharing over-represented features. Among these features, the biological processes (BP), molecular functions (MF) and cellular components (CC) annotating each gene are represented using the Gene Ontology (GO) [17]. GO is species-independent, and thus supports cross-species comparison [18]. The GO graph itself is also widely used for genes semantic similarity analysis [19].

Semantic similarity

Within a given gene set, the genes sharing identical or similar GO annotations can be grouped into clusters using two approaches [20]. The GSEA approach computes these clusters considering the GO terms over-representation. The semantic similarity approach takes into account GO properties to cluster genes considering the quantity and the importance of their shared annotations [21–24]. Both approaches are not exclusive, as semantic measures can be involved in GSEA in order to improve the analysis [25]. If these terms were independent, the gene set characterization could be performed by a straightforward set-based approach such as the Jaccard index or Dice's coefficient. However, GO terms are hierarchically-linked. Consequently, the characterization needs to take into account the underlying ontological structure of the GO annotations [26].

Semantic similarity measures rely on ontologies to systematically quantify the weight of the shared elements. They exploit the formal representation of the meaning of the terms by considering the relations between the terms (e.g. for inferring new annotations that were implicit as each term inherits all the properties of its ancestors) and by attributing different weights to each term depending on how much information they convey. When working with annotation databases, it should be routine practice to use the ontology hierarchy to infer implicit annotation [26]. Pesquita et al. performed an extensive review of the main semantic similarity measures [27] and identified two main categories, i.e. node-based methods and edge-based methods, as well as a handful of hybrid methods.

Node-based semantic similarity measures rely on how informative the terms are. Typically, they consider that two terms sharing an informative lowest common ancestor are more similar than two terms with a less informative lowest common ancestor. Historically, Information Content (IC) value was used to quantify how informative a term is, with the least frequent terms having the highest IC value. This concept, borrowed from Shannon's Information Theory [28], was used to measure similarities using ontologies [29–31] such as WordNet [32]. To compare two terms, these methods rely on their most informative common ancestor (MICA). The IC of this ancestor is the semantic similarity value between the compared terms. These methods developed in linguistics have been applied to GO [33, 34] using the frequency with which a term annotates a gene as a marker of its rarity. Consequently, the IC of a GO term is inversely proportional to the frequency with which it annotates a gene using the Gene Ontology Annotations (GOA) database [35]. GOA specifies also how each annotation has be attributed through Evidence Codes (EC). In their method called "IntelliGO", Benabderrahmane et al. use a weighting corresponding to each GO term EC in addition to their IC [36]. Retrieving only the most informative common ancestor to compute a semantic similarity ignores the possibility that two GO terms can share several common ancestors. These situations result in a loss of information. A possible solution has been proposed that consists in using the average of the IC values of all disjoint common ancestors (DCA) instead of the maximum IC of this common set [37]. For the node-based methods relying on IC, the terms' frequencies used to compute the IC values depend on the corpus of reference. In the context of genes comparison, IC-based methods have three main limits related to their dependence on a GOA-based corpus. First, it can prove difficult or even impossible to obtain a relevant corpus. GOA provides single and multi-species annotation tables. Although using a species-specific table is well-suited to intra-species comparisons, it becomes problematic for cross-species comparisons. Second, using a multi-species table (like the UniprotKB table) in these cases is biased towards the most extensively annotated species such as human or mice. Third, the well-studied areas of biology have high annotation frequencies and are therefore less informative and see their importance downgraded, whereas the less-studied areas are artificially upgraded [38–40].

Edge-based semantic similarity measures use the directed graph topology to compute distances between the terms to compare. Rada distance is based on the shortest path between the two terms [41]. Such distances rely on the average path among multiple paths [27]. Other approaches take into account the length of the path between the root of the ontology and the least common ancestor (LCA) of the terms, with the result that terms with a deep common ancestor are more similar than terms with a

common ancestor close to the root [42–46]. The edge-based methods using depth as a proxy for precision are not dependent on a particular corpus. This can be a good thing when it is difficult or impossible to determine a representative corpus, or a bad thing when corpus-dependent frequencies are relevant. Moreover, another constraint to consider is that granularity is not uniform in GO, so terms at the same depth can have different precisions [47].

Pesquita et al. also identified "hybrid" methods that combine different aspects of node-based and edge-based methods. In Wang's method [22], each term has a "semantic value" that represents how informative the term is, conforming to the node-based approach. However, the semantic value of a term is obtained by following the path from this term to the root and summing the semantic contributions of all the ancestors of this term. As the semantic value depends on the ontology topology, it also conforms to the edge-based approach.

Pesquita et al. do not single out any particular semantic similarity measure as the best one, as the optimal measure will depend on the data to compare and the level of detail expected in the results. The main advantage of Wang's method compared to purely node-based methods is that the semantic value is not GOA-dependent, unlike information content. It is thus well-suited to cross-species comparisons. As cross-species comparison is one of the key stakes in biology, further development in the domain of semantic comparison should support such comparisons.

Limitations of semantic similarity

All the semantic similarity measures appear appropriate for identifying and quantifying common features. However, as these measures are focusing on common features, they may lead to an incomplete analysis when comparing genes having particular features along side similar ones [48]. For example, parts A and B of Figure 1 respectively present the MF terms annotating the Exportin-5 orthologs of human (hsa) and rat (rno) and the Exportin-5 orthologs of human and drosophila (dme). Wang's method allows to compute cross-species semantic similarity. The results on MF annotations are: Sim(hsa, rno) = 0.797 and Sim(hsa, dme) = 0.726. This is consistent with the fact that globally, the Exportin-5 orthologs share the same functions between hsa, rno and dme. However, there are also five times as many human-specific MF terms compared to drosophila as compared to rats. It has been demonstrated that Exportin-5 orthologs are functionally divergent among species [49]. The tiny difference of semantic similarity (0.071) correctly reflects the fact that the orthologs share the same main function, but is not sufficient to identify that some species also have additional functions.

We assume that considering only similarity measures is not enough to compare sets of annotations. This analysis is valid for any set of annotations that refer to an ontology. We hypothesize that gene set analysis can be improved by considering gene particularities in addition to gene similarities. We propose a general definition and some associated formal properties. We propose also a new approach based on the notion of GO term informativeness to compute gene set particularities.

Method

Definition of semantic particularity

The semantic particularity of a set compared to another is the value that reflects the importance of the features that belong to the first set but not the second. To compare two genes, we rely on the similarity and the respective particularities of their sets of annotations. The particularity of a gene g1 annotated by the set Sg1 compared to a gene g2 annotated by the set Sg2 depends on the annotations of Sg1 that are not related to any annotation of Sg2.

Formal properties

Like for semantic similarity, we compute a value bounded by 0 (least particular) and 1 (most particular). Four important properties arise from the semantic particularity definition:

- The semantic particularity is asymmetric:

$$\text{Par}(\text{Sg1}, \text{Sg2}) = \text{x} \nRightarrow \text{Par}(\text{Sg2}, \text{Sg1}) = \text{x} \qquad \textbf{(Prop 1)}$$

- Compared to itself, a set of annotations has no semantic particularity:

$$\text{Par}(\text{Sg1}, \text{Sg1}) = 0 \qquad \textbf{(Prop 2)}$$

  If $\text{Sg1} = \varnothing$, this comparison is meaningless.

- The semantic particularity of a set of annotations Sg1 ($\neq \varnothing$) is maximal when it is compared to an empty set of annotations:

$$\text{Par}(\text{Sg1}, \varnothing) = 1 \qquad \textbf{(Prop 3.1)}$$

  And conversely:

$$\text{Par}(\varnothing, \text{Sg1}) = 0 \qquad \textbf{(Prop 3.2)}$$

- The particularity of a set Sg1 of annotations compared to a set Sg2 does not depend on the elements of Sg2 that do not belong to Sg1:

$$\text{Sg3} \cap \text{Sg1} = \varnothing \Rightarrow \text{Par}(\text{Sg1}, \text{Sg2}) = \text{Par}(\text{Sg1}, \text{Sg2} \cup \text{Sg3}) \qquad \textbf{(Prop 4)}$$

Measure of semantic particularity

In order to compute the particularity of Sg1 compared to Sg2, we focus on the terms of Sg1 that are not members of Sg2. This requires to address two problems: the terms are not independent, and they do not convey the same amount of information.

Some of the terms of Sg1 that are not members of Sg2 may be linked in the graph. Taking several linked terms into account would result in considering them several times. For example, in Figure 1B, considering both "RNA binding" and "tRNA binding" would result in counting twice the contribution of "RNA binding". Therefore, we should only focus on the terms of Sg1 that do not have any descendant in Sg1 and that are not members of Sg2. Some of these terms might be ancestors of terms of Sg2 and should be considered as common to Sg1 and Sg2. We call Sg^* the union of Sg and the sets of ancestors of each element of Sg. We call MPT(Sg1, Sg2) the set of most particular terms of Sg1 compared to Sg2. MPT(Sg1, Sg2) is the set of terms of Sg1 that do not have any descendant in Sg1 and that are not members of Sg2^*. In the Figure 1B, MPT(hsa, dme) = ["tRNA binding"].

Using the set theory, we could define Par(Sg1, Sg2) as the proportion of elements of Sg1 that belong to MPT(Sg1, Sg2). When computing card(MPT(Sg1, Sg2)), all the elements have the same weight. However, considering the semantics underlying these elements, some of them may be more informative than others and should ideally be emphasized. Different strategies, similar to those already proposed for the computation of the semantic similarity, can be applied.

We then define PI(Sg1, Sg2), the particular informativeness of a set of GO terms Sg1 compared to another set of GO terms Sg2, as the sum of the differences between the informativeness (I) of each term

t_p of MPT(Sg1, Sg2) and the informativeness of the most informative common ancestor (MICA) between t_p and Sg2. The PI of a set of terms is the information that is not shared with the other set.

$$PI(Sg1, Sg2) = \sum_{t_p \in MPT(Sg1,Sg2)} I(t_p) - I(MICA(t_p, Sg2)) \quad (1)$$

In the Figure 1B, PI(hsa, dme) = I(tRNA binding) - I(binding). We have no sum in this example since MPT(Sg1, Sg2) only contains one term.

We last normalize PI to compute Par(Sg1, Sg2), the semantic particularity of the set of GO terms Sg1 compared to the set of GO terms Sg2. We define MCT(Sg1, Sg2), the set of the most informative common terms of Sg1 and Sg2, as the set of the terms belonging to the intersection of Sg1* and Sg2* that do not have any descendant either in Sg1* or in Sg2*. In the Figure 1B, MCT(hsa, dme) = ["protein transporter activity", "protein binding"]. Par(Sg1, Sg2) is the ratio of PI(Sg1, Sg2) and the sum of the informativeness of Sg1 most informative terms (i.e. those Sg1-specific and those common with Sg2; the MICA in the PI formula for the Sg1-specific guarantees that the informativeness of common terms is not counted twice).

$$Par(Sg1, Sg2) = \frac{PI(Sg1, Sg2)}{PI(Sg1, Sg2) + \sum_{t_c \in MCT(Sg1,Sg2)} I(t_c)} \quad (2)$$

For the example of the Figure 1B, this formula becomes:

$$Par(hsa, dme) = \frac{I(tRNA\ binding) - I(binding)}{(I(tRNA\ binding) - I(binding)) + (I(p.\ trsp.\ activity) + I(protein\ binding))} \quad (3)$$

Several measures of informativeness have been proposed. The widely used Information Content (IC) family depends on an annotation corpus (e.g. GOA). The IC of a term t is its negative log probability $P(t)$.

$$IC(t) = -log(P(t))$$

In the context of GO terms comparison, the probability of occurrence of a term $P(t)$ is estimated by its frequency in annotations [27]. It is necessary to take into account Gene Ontology subsumption hierarchy when computing this frequency in order to also consider implicit annotations to the terms descendants [26]. IC is typically used when a representative corpus is available such as human GOA for studying human genes functions.

The alternative approach is corpus-independent. A term informativeness is a function of its distance to the root. It is typically used when a relevant corpus cannot be computed (for comparing elements from several species) or does not exist (for poorly studied species). Wang's Semantic Value (SV) computes this type of informativeness. The relevance of the results obtained by this approach has previously been demonstrated [22, 27]. Wang first computes the semantic contributions of the ancestors of each term to compare to these terms, following:

$$\begin{cases} S_A(A) = 1 \\ S_A(t) = max\{w_e * S_A(t') \mid t' \in children\ of\ (t)\}\ if\ t \neq A \end{cases}$$

where $S_A(t)$ is the semantic contribution of the term t to the term A and w_e is the semantic contribution factor for edge e linking a term t with its child term t'. According to Wang, we use a semantic contribution factor of 0.8 for the "is a" relations and 0.6 for the "part of" relations, and we added a 0.7 factor for the "[positively] [negatively] regulates" relations. An additional study not presented here showed that the value of the regulation factor had minimal impact (+/- 0.01) on the overall value.

Then, for each target term to compare, the semantic value is the sum of the semantic contributions of all its ancestors:

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

As shown in the equation 3, four terms are involved in the calculation of the MF particularity of the human Exportin-5 ortholog compared to the drosophila Exportin-5 ortholog. This comparison is cross-species, so a semantic value-based informativeness measure is relevant. According to the previous formula, the semantic values of the terms involved in the equation 3 are: SV(tRNA binding) = 4.201, SV(binding) = 1.8, SV(protein transporter activity) = 2.952 and SV(protein binding) = 2.44. Consequently, we can compute: Par(hsa, dme) = 0.308. Likewise, for Figure 1A, Par(hsa, rno) = 0.082.

Results

To study the benefits of our approach over an analysis based only on similarity, we considered three biological cases. In order to determine if we could extend Wang's initial results, our first use case was *Saccharomyces cerevisiae* tryptophan degradation. As both the ontology and the annotations have evolved since 2007 [39], we computed the updated semantic similarity. Then, we computed the particularity measure in order to evaluate its benefits. In case 2, we computed the similarity and particularity values on a set of 51 gene products belonging to a same human metabolic pathway. The motivation is to study whether the results of the case 1 can be generalized to a larger set of genes. We also studied how using IC-based or semantic value-based similarity and particularity measures affects the conclusions. In case 3, we applied the semantic similarity and particularity measures on all the groups of homolog genes from the the HomoloGene database. This approach aims to identify systematically homologues expected to be similar and having also particular functions.

In all these cases, we used the GOSemSim R package to compute Lin's similarity and to provide IC tables used in the computation of the IC-based particularity [50]. We used a personal implementation of Wang's similarity and the corresponding SV used in SV-based particularity computation.

Case 1: *Saccharomyces cerevisiae* tryptophan degradation

We first tested our approach on the example chosen by Wang [22]: *Saccharomyces Cerevisiae* tryptophan degradation [51]. We computed the semantic similarity according to Wang's method (Table 1) using the most recent version of annotation data available (August 2013 versions of GOA and GO).

Wang's conclusions remained true: we can still distinguish the three groups of genes involved in the three main steps of tryptophan degradation. Similarity values for the group [ARO8, ARO9] involved in the first step were 0.92. Similar results were observed for the group [ARO10, PDC6, PDC5, PDC1] involved in the second step and for the group [SFA1, ADH5, ADH4, ADH3, ADH2, ADH1] involved in the last step. The similarities measured between genes of 2 different groups ("inter-group measures") were greater than in Wang's original study but remained lower than the intra-group comparison measures. We found the same three groups as Wang. These groups are biologically relevant because they are involved in the three steps of *Saccharomyces cerevisiae* tryptophan degradation pathway. To obtain these groups, Wang used a threshold of 0.770 in 2007. We used a threshold of 0.745.

We completed the previous results with the measures of semantic particularity, using Wang's Semantic Value as informativeness (Table 2). The highest particularity values were between genes from different groups which is consistent with the analysis of the semantic similarity values.

Our approach also identified a characteristic of the compared genes that the similarity ignored. Indeed, some of the genes belonging to the same group have also some particular functions (i.e. high similarity and relatively high particularity). For example, all the genes of the third group are similar. However,

Table 2 shows that all the genes of this group have a high particularity value compared to ADH4. Notably, the similarity between SFA1 and ADH4 was 0.745 and SFA1 particularity was 0.388 whereas most of the other intra-group particularity values in this group were zero or close to zero. Figure 2 presents the distribution of GO annotations between genes ADH4 and SFA1. It shows that the observed particularity value is mostly related to SFA1-specific nucleotide binding function. So, two genes can be similar while at least one of them has some particular functions.

The similarity values show that Wang results are still valid. We also identified a benefit of using a particularity measure in addition to a similarity measure for identifying particular functions between similar genes.

Case 2: *Homo sapiens* aquaporin-mediated transport

In the previous case, we found an example of a relatively high particularity value between similar genes. In this second case, we aim to study a larger dataset in order to determine the frequency and the importance of this situation. We used a dataset composed by 51 well-annotated human genes involved in the aquaporin-mediated transport pathway for *Homo sapiens*. We used the list of all involved genes provided by the Reactome database [52]. In continuity with the first case, we computed the Wang similarity and S-Value-based particularities for each pair of genes of this list. As the Human annotation database is one of the most comprehensive, we also duplicating the study using Lin's measure as an IC-based similarity, and IC as a value of GO term informativeness for our specificity. All the results are available in supplementary file 1. Table 3, 4 and 5 present the average, standard deviation, minimum and maximum values of particularity measured in this study for each branch of GO. We classified these statistics in 20 similarity categories containing all the comparison results ranging from sim = 0.5 to sim = 0.999 with steps of sim = 0.025.

The relatively high particularity between similar genes that we observed in case 1 is confirmed in this case 2. In each 20 categories in the human aquaporin-mediated transport pathway, some of the genes have an important particularity compared to the others. Again, these genes cannot be identified using only a similarity measure.

Figure 3 illustrates this case giving the MF annotation graph of two couples of genes: AQP8 and AQP5 in part A and AQP6 and AQP3 in part B. The corresponding similarity and particularity values are presented in table 6. The two couples have close similarity values regardless the method used but they show a very different particularity profile, with much higher particularities between AQP6 and AQP3 than between AQP8 and AQP5. The two distinct informativeness measures used to compute the particularity led to the same conclusion. The same phenomenon can be observed in the 20 categories of similar genes.

These results confirm that among similar genes, some also have some particular functions, and show that this situation can be observed throughout the full range of similarity values. Therefore, the situation described in the first use case was not an isolated case.

Case 3: Homologs comparison

The previous cases focused on the similarity and particularity of different genes in a same pathway. In this third case, we compared homolog genes across different species. IC-based methods cannot be used in this cross-species context. To investigate the frequency of similar homolog genes and the frequency of homolog genes having particular functions, we computed Wang's semantic similarity and SV-based particularities for each group of the HomoloGene database. The August 2013 version of this database contained 43,074 groups of homolog genes. Each group contained from 2 to 839 genes (average: 6.02, standard deviation: 7.46). We computed all the 5,531,994 intra-group similarity and particularity measures. Table 7 categorizes the comparisons according to the number of annotated genes.

To be valid, a comparison has to involve two annotated genes. Overall, 21.94% of the comparisons were valid. For BP, CC and MF, we used the number of valid comparisons as the baseline to analyze the different configurations of similarity and particularity. We focused on these valid comparisons and found that 89.93% of them had a similarity greater than or equal to 0.5. In 82.26%, the genes were similar and had particularities lower than 0.5. Although there were differences between BP, MF and CC, on the whole HomoloGene database, the particularity values allowed us to identify 7.63% of the valid comparisons that denote similar genes, one of these genes having a particularity greater than 0.5.

As an example illustrating the results, we analyzed the comparisons of the GO molecular functions associated to Exportin-5 orthologs for 9 species (table 8). 27 of the 36 comparisons (75%) involved pairs of genes with a similarity greater than 0.5. 12 of these 27 comparisons involved similar pairs of genes, one of them having a particularity greater than 0.3 (mostly for *Canis canis* and *Drosophila melanogaster*). Among these, five comparisons involving *Canis canis* resulted in a similarity value over 0.5 and one particularity value over 0.5. The remaining 9 of the 36 comparisons involved genes with a similarity lower than 0.5 and particularities greater than 0.5 (mostly for *Arabidopsis thaliana* and one for *Canis canis*).

Altogether, the case 3 results showed that ortholog genes were, as expected, mostly similar. We have also demonstrated that some of them may have high particularity values that denote particular functions. Last, some orthologs may have diverged to present a low similarity and high particularities.

Discussion

Semantic particularity

Semantic similarity measures have been extensively used for comparing genes and gene sets [19] but they only tell a part of the story. Similarity is symmetric. It decreases slowly as the number of gene-particular annotations increases. However, similarity alone does not indicate which gene has some particular functions and does not even reveal these particular functions. There is a need for a measure to qualify this particularity (does gene1 have some particular functions compared to gene2, even if gene1 and gene2 are similar?) and to quantify these respective differences (what is the importance of gene1's particular functions compared to gene2?). Simple comparisons of the sets of terms annotating two genes, such as Venn diagram representations, give an initial picture of each gene's particularity. However, this approach is biased due to the relations between the terms of an ontology. Like for similarity, measuring particularity has to take semantics into account. Diaz-Diaz et al. proposed a semantic approach to compute a dissimilarity measure in order to evaluate the functional coherence of entire gene sets [46]. The dissimilarity of two terms is obtained by measuring a distance in edges in the GO graph and weighting the result with the depth of the considered terms, as in Wu and Palmer's similarity measure [43]. This notion of dissimilarity is therefore strongly related to similarity and does not provide a way to compute the particularity as we defined earlier (high dissimilarity indicates low similarity, and vice versa). However, the two categories of similarity measures, i.e. "edge-based" and "node-based", can be used for this purpose. Each approach has its drawbacks [27]. Edge-based methods are biased because the GO terms are not homogeneously distributed across the tree, while node-based methods that use an IC value are dependent on a specific annotation corpus, which puts a limit on their use for cross-species comparisons. In cross-species studies, it is impossible to compare IC values relying on term frequencies obtained from different corpora. Using a global corpus instead, such as the UniprotKB GOA table is biased in favor of the most studied functions in the most studied species. Therefore, graph-based approaches relying on the distance to the root are more appropriate in such situation.

We based our semantic particularity measure on the concept of informativeness of GO terms. This informativeness can either be an Information Content (IC) [29–31, 33, 34] value or a Semantic Value (SV) [22]. The choice between these two alternatives depends on the data to compare. IC is preferred to compare genes from a same species when an important annotation corpus is available for this species. SV

is preferred to compare genes from different species or genes from a same species without an important annotation corpus. Therefore, we advise to use a combination of either IC-based or of SV-based similarity and particularity measures when computing profiles based on similarity and particularity values.

The interpretation of the similarity and particularity values depends on the number and quality of the annotations. If at least one of two genes has few annotations, the similarity and particularity values will suffer from a lack of precision (the values are sensitive to the addition of new annotations) regardless of their accuracy.

Furthermore, annotations are associated with different Evidence Codes (EC), ranging from automatic inference to experimental validation. The biological interpretation of similarity and particularity values is more convincing when their computation refers to experimentally-confirmed annotations. However, electronically-inferred annotations may still yield valid similarity and particularity values. As the GO consortium recommends against using EC as a measure of quality of the annotation [53], we did not use them to weight the similarity and particularity values. However, we paid attention to this aspect when interpreting the results of our case studies. Our approach consisted in comparing two genes using a tuple of one symmetric similarity value and the two particularity values. Having high similarity and low particularities for two genes indicates that these genes globally have the same characteristics in the compared domain (BP, MF or CC) and none of them has any major additional particularity. Conversely, a low similarity and high particularities between two genes indicates that these genes are different in the compared domain. Furthermore, among highly similar genes, finding that one gene has also a high particularity value allows to identify additional features for this gene not present in the other one despite their high similarity. This contributed to a more accurate analysis than using similarity alone by distinguishing interesting sub-groups of features with close similarity values.

Case studies: benefits of the semantic particularity

Particularity refined the similarity-based analysis by identifying some couples of similar genes with important particularities. All three use cases illustrated this point in intra-species or in cross-species.

In the first case study on the *Saccharomyces Cerevisiae* tryptophan degradation pathway, SFA1 and ADH4 had similarity values close to those of the other genes of the same sub-group. However, SFA1 and to a lesser extent all the other genes that catalyze the same reaction had some particular functions compared to ADH4. Consequently, it is possible that two similar genes also have some particular functions (i.e. high similarity and relatively high particularity). The particularity is not systematically inversely proportional to the similarity. Moreover, some of these these atypical cases may be of biological interest.

We have gone further in the case 2, comparing 51 genes that belong to a same human pathway. With this case, we wanted to see three things. First, we wanted to know whether the observations made in the first case remained true on a bigger example. They did. Then, we wanted to assess the effect of the kind of informativeness used. Semantic value and information content gave different semantic similarity and particularity values, but they leaded to the same conclusions. Consequently, the choice of this method only depends on the data we want to compare. IC can be used as an informativeness measure if the data are relative to one single species and if this species is sufficiently annotated to offer a meaningful corpus. Otherwise, the best informativeness measure may be the semantic value. Last, we wanted to assess our conclusions on the three branches of Gene Ontology. Concerning this point, we obtained high particularity values between similar genes regarding any branch of GO.

The third case showed comparisons of ortholog genes that also resulted in interesting sub-cases with high-similarity profiles. As suspected, the results confirmed that ortholog genes are mostly similar. Moreover, particularity measures made it possible to observe that among the pairs of similar genes, some are composed of at least one gene having also an important particularity. Indeed, among the 1,213,588 valid comparisons across the whole HomoloGene database, we identified 93,152 (7.68%) comparisons for which the genes were similar, but at least one of them had an important particularity, denoting some particular function(s). This confirm the observations made in the cases 1 and 2. These 7.68% of valid

comparisons resulting in the identification of genes having some particular features, which however have enough common GO annotation to remain similar are biologically very interesting. This demonstrates the benefits of using the semantic particularity measure in addition to semantic similarity.

In the third case, we developed the Exportin-5 example to illustrate the limitations of the semantic similarity measures. The results of a similarity measure did not reflect that the amount of particular functions while comparing the human gene to the drosophila ortholog ("tRNA binding" and four of its ancestors are human-specific) is greater than while comparing it to the rat ortholog (only "protein binding" is human-specific). The particularity measure showed that the human and drosophila Exportin-5 orthologs are not only similar, but that some quantifiable features are in reality very specific to the human gene. Furthermore, the high particularity of these orthologs is consistent with the results of Shibata et al., who demonstrated that Exportin-5 orthologs are functionally divergent among species [49].

Interpretation of similarity and particularity values

The case studies showed that combining similarity and particularity makes it possible to identify some genes' particular functions that cannot be distinguished using similarity only. These particular functions may be the result of a real biological difference, a default of annotation, or a combination of both. If we suspect a default of annotation, the results should be interpreted carefully until the annotations are improved.

In the case 3, the number of annotations vary between the compared orthologs. On the one hand, the results can reflect a real particularity of function for some genes. On the other hand, the high particularity of a gene can be the result of a lack of annotations of the other gene. For example, when comparing MF annotations for hsa and ath orthologs of Exportin-5, we observed very high particularities for both species (respectively 0.641 and 0.871). We consider these results to be relevant, as the genes of both species are well annotated (11 annotations in the expanded set of hsa, 18 annotations in the expanded set of ath). Conversely, care is warranted when interpreting the particularity of hsa over Canis canis (cca). For these species, sim(hsa, cca) = 0.428, spe(hsa, cca) = 0.611 and spe(cca, hsa) = 0. However, the expanded set of annotations for the cca ortholog had only 4 terms compared to 11 for hsa. In this case, the high particularity of hsa could be attributed to the lack of cca annotations.

Synthesis

We showed that gene set analysis can be improved by considering gene-set particularities in addition to their similarity. We proposed a set of formal properties and a new GO semantic measure to compute gene-set particularity. We first showed that particularity is a useful complement to similarity for comparing gene sets, making it possible to detect similar gene sets for which one of the sets also had some particular functions, and to identify these functions. We also showed that using particularity also improves gene clustering. Our particularity measure relies on the informativeness of GO terms. This informativeness of a term can be its Information Content or its Semantic Value. In this paper, we combined our particularity measure with a similarity measure to compare genes annotated GO terms, but this same principle can be generalized to other ontologies.

Acknowledgments

CB was supported by a fellowship from the French ministry of research.

References

1. Cannata N, Merelli E, Altman RB (2005) Time to organize the bioinformatics resourceome. PLoS Comput Biol 1: e76.

2. Grossmann S, Bauer S, Robinson PN, Vingron M (2007) Improved detection of overrepresentation of gene-ontology annotations with parent child analysis. Bioinformatics 23: 3024–31.

3. Klie S, Mutwil M, Persson S, Nikoloski Z (2012) Inferring gene functions through dissection of relevance networks: interleaving the intra- and inter-species views. Mol Biosyst 8: 2233–41.

4. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1–13.

5. Barriot R, Sherman DJ, Dutour I (2007) How to decide which are the most pertinent overly-represented features during gene set enrichment analysis. BMC Bioinformatics 8: 332.

6. Stobbe MD, Jansen GA, Moerland PD, van Kampen AHC (2012) Knowledge representation in metabolic pathway databases. Brief Bioinform .

7. Hawkins T, Chitale M, Kihara D (2010) Functional enrichment analyses and construction of functional similarity networks with high confidence function prediction by pfp. BMC Bioinformatics 11: 265.

8. Teng Z, Guo M, Liu X, Dai Q, Wang C, et al. (2013) Measuring gene functional similarity based on group-wise comparison of go terms. Bioinformatics 29: 1424–32.

9. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–50.

10. Maere S, Heymans K, Kuiper M (2005) Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21: 3448–9.

11. Zheng Q, Wang XJ (2008) Goeast: a web-based software toolkit for gene ontology enrichment analysis. Nucleic Acids Res 36: W358–63.

12. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, et al. (2009) Cluego: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 25: 1091–3.

13. Sherman BT, Huang DW, Tan Q, Guo Y, Bour S, et al. (2007) David knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics 8: 426.

14. Baker EJ, Jay JJ, Bubier JA, Langston MA, Chesler EJ (2012) Geneweaver: a web-based system for integrative functional genomics. Nucleic Acids Res 40: D1067–76.

15. Zhang B, Schmoyer D, Kirov S, Snoddy J (2004) Gotree machine (gotm): a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies. BMC Bioinformatics 5: 16.

16. Hung JH, Yang TH, Hu Z, Weng Z, DeLisi C (2012) Gene set enrichment analysis: performance evaluation and usage guidelines. Brief Bioinform 13: 281–91.

17. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. Nat Genet 25: 25–9.

18. Primmer CR, Papakostas S, Leder EH, Davis MJ, Ragan MA (2013) Annotated genes and nonannotated genomes: cross-species use of gene ontology in ecology and evolution research. Mol Ecol 22: 3216–3241.

19. Wang L, Jia P, Wolfinger RD, Chen X, Zhao Z (2011) Gene set analysis of genome-wide association studies: methodological issues and perspectives. Genomics 98: 1–8.

20. Ochs MF, Peterson AJ, Kossenkov A, Bidaut G (2007) Incorporation of gene ontology annotations to enhance microarray data analysis. Methods Mol Biol 377: 243–54.

21. Ovaska K, Laakso M, Hautaniemi S (2008) Fast gene ontology based clustering for microarray experiments. BioData Min 1: 11.

22. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF (2007) A new method to measure the semantic similarity of go terms. Bioinformatics 23: 1274–81.

23. Kustra R, Zagdanski A (2006) Incorporating gene ontology in clustering gene expression data. In: CBMS. IEEE Computer Society, pp. 555–563.

24. Bolshakova N, Azuaje F, Cunningham P (2005) A knowledge-driven approach to cluster validity assessment. Bioinformatics 21: 2546–7.

25. Chang B, Kustra R, Tian W (2013) Functional-network-based gene set analysis using gene-ontology. PLoS One 8: e55635.

26. Rhee SY, Wood V, Dolinski K, Draghici S (2008) Use and misuse of the gene ontology annotations. Nat Rev Genet 9: 509–15.

27. Pesquita C, Faria D, Falcão AO, Lord P, Couto FM (2009) Semantic similarity in biomedical ontologies. PLoS Comput Biol 5: e1000443.

28. Shannon CE (1948) A mathematical theory of communication. Bell system technical journal 27.

29. Resnik P (1999) Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence 11: 95–130.

30. Lin D (1998) An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning : 296–304.

31. Jiang J, Conrath D (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the International Conference Research on Computational Linguistics (ROCLING). Taiwan.

32. Miller G (1995) Wordnet: A lexical database for english. Communications of the ACM 38: 39–41.

33. Lord PW, Stevens RD, Brass A, Goble CA (2003) Semantic similarity measures as tools for exploring the gene ontology. In: Pacific Symposium on Biocomputing. pp. 601–612.

34. Sheehan B, Quigley A, Gaudin B, Dobson S (2008) A relation based measure of semantic similarity for gene ontology annotations. BMC Bioinformatics 9: 468.

35. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, et al. (2004) The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. Nucleic Acids Res 32: D262–6.

36. Benabderrahmane S, Smail-Tabbone M, Poch O, Napoli A, Devignes MD (2010) Intelligo: a new vector-based semantic similarity measure including annotation origin. BMC Bioinformatics 11: 588.

37. Couto FM, Silva MJ, Coutinho PM (2007) Measuring semantic similarity between gene ontology terms. Data & Knowledge Engineering 61: 137–152.

38. Jin B, Lu X (2010) Identifying informative subsets of the gene ontology with information bottleneck methods. Bioinformatics 26: 2445–51.

39. Gillis J, Pavlidis P (2013) Assessing identity, redundancy and confounds in gene ontology annotations over time. Bioinformatics 29: 476–82.

40. Chen G, Li J, Wang J (2013) Evaluation of gene ontology semantic similarities on protein interaction datasets. Int J Bioinform Res Appl 9: 173–83.

41. Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man, and Cybernetics 19: 17–30.

42. Pekar V, Staab S (2002) Taxonomy learning - factoring the structure of a taxonomy into a semantic classification decision. In: COLING.

43. Wu Z, Palmer M (1994) Verb semantics and lexical selection. In: Proc. of the 32nd annual meeting on Association for Computational Linguistics. pp. 133–138. doi: http://dx.doi.org/10.3115/981732.981751.

44. Cheng J, Cline M, Martin J, Finkelstein D, Awad T, et al. (2004) A knowledge-based clustering algorithm driven by gene ontology. J Biopharm Stat 14: 687–700.

45. Alvarez MA, Yan C (2011) A graph-based semantic similarity measure for the gene ontology. J Bioinform Comput Biol 9: 681–95.

46. Díaz-Díaz N, Aguilar-Ruiz JS (2011) Go-based functional dissimilarity of gene sets. BMC Bioinformatics 12: 360.

47. Mazandu GK, Mulder NJ (2012) A topology-based metric for measuring term similarity in the gene ontology. Adv Bioinformatics 2012: 975783.

48. Clark WT, Radivojac P (2013) Information-theoretic evaluation of predicted ontological annotations. Bioinformatics 29: i53–61.

49. Shibata S, Sasaki M, Miki T, Shimamoto A, Furuichi Y, et al. (2006) Exportin-5 orthologues are functionally divergent among species. Nucleic Acids Res 34: 4711–21.

50. Yu G, Li F, Qin Y, Bo X, Wu Y, et al. (2010) Gosemsim: an r package for measuring semantic similarity among go terms and gene products. Bioinformatics 26: 976–8.

51. Saccharomyces cerevisiae tryptophan degradation pathway from yeastcyc website. Available: `http://goo.gl/uKGiRH`.

52. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, et al. (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 39: D691–7.

53. Guide to go evidence codes of gene ontology website. Available: `http://goo.gl/LUBrb`.

Figures

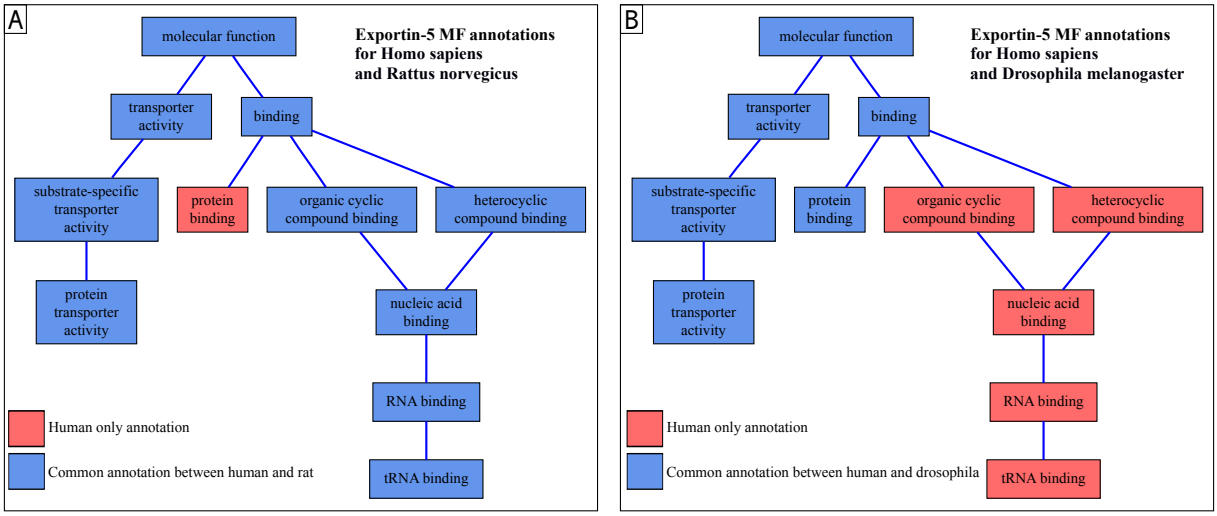


Figure 1. Representation of Exportin-5 orthologs annotations. Part A of this figure displays the MF annotations of the human and rat orthologs of Exportin-5. Part B displays the MF annotations of the human and drosophila orthologs of Exportin-5. Common terms between species are displayed in blue. The terms annotating only the human ortholog are displayed in red. In this example, there is no rat nor drosophila-specific term. The semantic similarity values obtained in these cases do not reflect the difference of human particularity between each part.

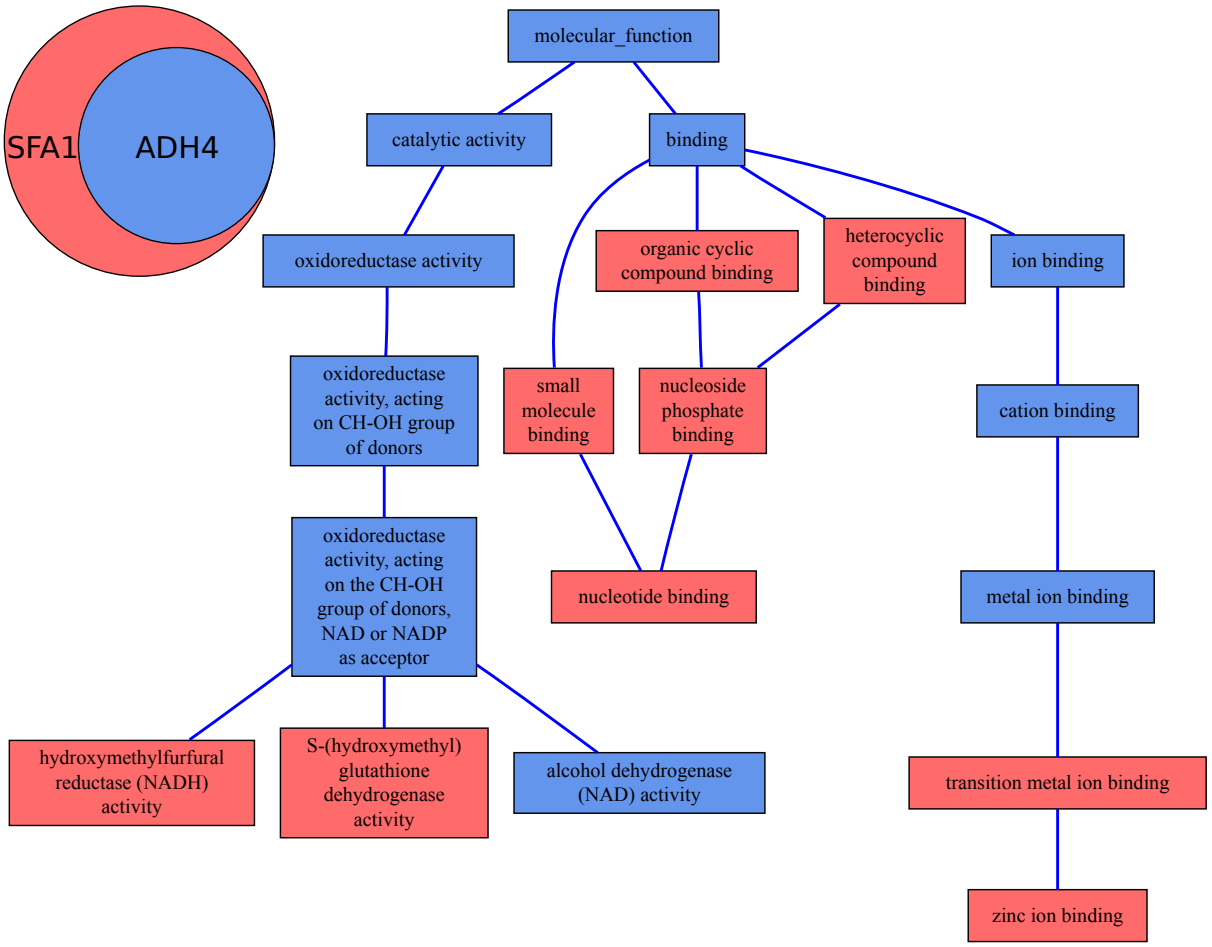


Figure 2. Representation of ADH4 and SFA1 *Saccharomyces cerevisiae* annotations. The particularity of 0.388 for SFA1 compared to ADH4 is explained notably by the term "nucleotide binding", to which the closest ancestor with ADH4 annotations is at a distance of three edges. The other red terms are also responsible for this particularity.

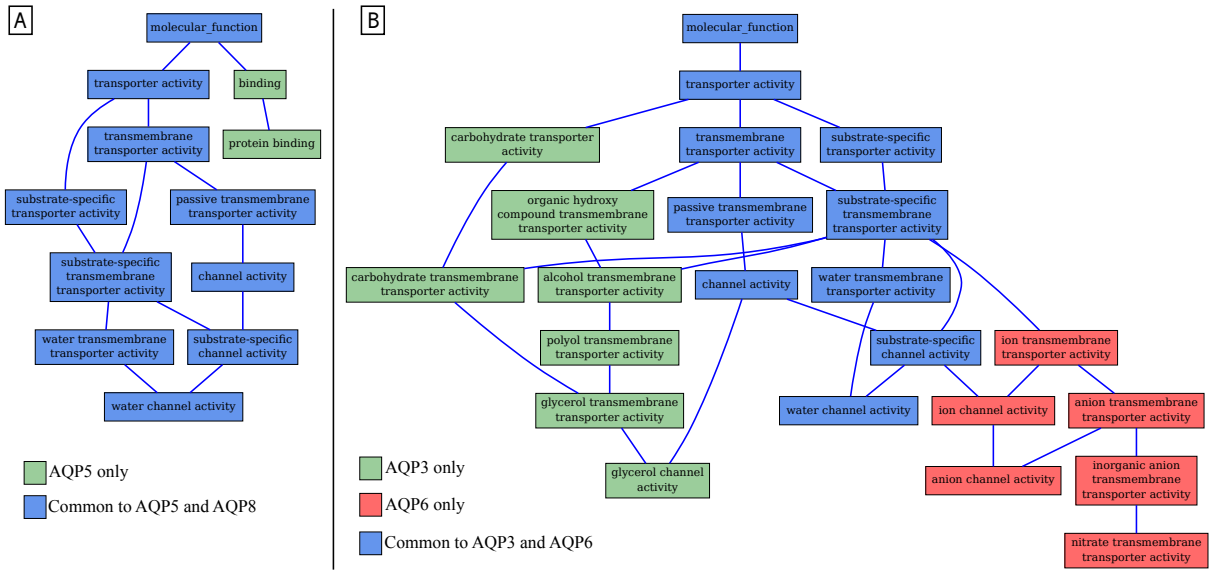


Figure 3. MF annotations of two couples of human aquaporins. Part A: AQP8 and AQP5 share most of their annotations. Part B: AQP6 and AQP3 share numerous molecular functions, but each gene also have particular functions.

Tables

Table 1. Semantic similarity values between genes involved in the *Saccharomyces cerevisiae* tryptophan degradation pathway

MF			ARO9	ARO8	ARO10	PDC6	PDC5	PDC1	SFA1	ADH5	ADH4	ADH3	ADH2	ADH1
		Annots	15	14	21	20	20	20	19	17	10	17	17	18
SIM	ARO9	15	1	0.92	0.283	0.295	0.295	0.295	0.229	0.237	0.269	0.237	0.237	0.233
	ARO8	14		1	0.287	0.299	0.299	0.299	0.232	0.241	0.273	0.241	0.241	0.236
	ARO10	21			1	0.961	0.961	0.961	0.352	0.395	0.449	0.395	0.395	0.371
	PDC6	20				1	1	1	0.371	0.428	0.499	0.428	0.428	0.395
	PDC5	20					1	1	0.371	0.428	0.499	0.428	0.428	0.395
	PDC1	20						1	0.371	0.428	0.499	0.428	0.428	0.395
	SFA1	19							1	0.932	0.745	0.932	0.932	0.91
	ADH5	17								1	0.751	1	1	0.961
	ADH4	10									1	0.751	0.751	0.747
	ADH3	17										1	1	0.961
	ADH2	17											1	0.961
	ADH1	18												1

Color gradient according to similarity value (0 = white, 1 = blue). The given numbers of annotations ("Annots") consider the GO terms that annotate directly the genes and their ancestors.

Table 2. Semantic particularity values between genes involved in the *Saccharomyces cerevisiae* tryptophan degradation pathway

MF			ARO9	ARO8	ARO10	PDC6	PDC5	PDC1	SFA1	ADH5	ADH4	ADH3	ADH2	ADH1
		Annots	15	14	21	20	20	20	19	17	10	17	17	18
PAR	ARO9	15	0	0.047	0.515	0.515	0.515	0.515	0.658	0.658	0.658	0.658	0.658	0.658
	ARO8	14	0.026	0	0.506	0.506	0.506	0.506	0.65	0.65	0.65	0.65	0.65	0.65
	ARO10	21	0.62	0.62	0	0.019	0.019	0.019	0.464	0.464	0.677	0.464	0.464	0.464
	PDC6	20	0.578	0.578	0	0	0	0	0.417	0.417	0.634	0.417	0.417	0.417
	PDC5	20	0.578	0.578	0	0	0	0	0.417	0.417	0.634	0.417	0.417	0.417
	PDC1	20	0.578	0.578	0	0	0	0	0.417	0.417	0.634	0.417	0.417	0.417
	SFA1	19	0.711	0.711	0.415	0.415	0.415	0.415	0	0.049	0.388	0.049	0.049	0.049
	ADH5	17	0.61	0.61	0.289	0.289	0.289	0.289	0	0	0.351	0	0	0
	ADH4	10	0.399	0.399	0.268	0.268	0.268	0.268	0	0	0	0	0	0
	ADH3	17	0.61	0.61	0.289	0.289	0.289	0.289	0	0	0.351	0	0	0
	ADH2	17	0.61	0.61	0.289	0.289	0.289	0.289	0	0	0.351	0	0	0
	ADH1	18	0.668	0.668	0.358	0.358	0.358	0.358	0.025	0.025	0.37	0.025	0.025	0

Color gradient according to particularity value (0 = white, 1 = red or green). If Par(gene1, gene2) is displayed in green, Par(gene2, gene1) is displayed in red. The value contained in a cell is the particularity of the gene displayed at its row header compared to the gene displayed at its column header. For example, Par(ARO10, ARO8) = 0.62 and Par(ARO8, ARO10) = 0.506. The given numbers of annotations ("Annots") consider the GO terms that annotate directly the genes and their ancestors.

Table 3. Particularity value statistics in 20 similarity values ranges from case 2 - BP measures

BP	S-value-based particularity				IC-based particularity			
Similarity	Average	Std dev.	Min	Max	Average	Std dev.	Min	Max
[0.5-0.524]	0.401	0.2	0.013	0.844	0.562	0.223	0	0.904
[0.525-0.549]	0.386	0.174	0	0.794	0.532	0.284	0	0.89
[0.55-0.574]	0.347	0.199	0	0.707	0.497	0.244	0	0.886
[0.575-0.599]	0.352	0.198	0	0.798	0.502	0.241	0	0.895
[0.6-0.624]	0.315	0.203	0	0.671	0.495	0.208	0	0.794
[0.625-0.649]	0.292	0.145	0	0.629	0.437	0.25	0	0.882
[0.65-0.674]	0.299	0.162	0	0.615	0.439	0.258	0	0.876
[0.675-0.699]	0.229	0.15	0	0.529	0.451	0.216	0.039	0.839
[0.7-0.724]	0.228	0.166	0	0.631	0.403	0.239	0	0.859
[0.725-0.749]	0.22	0.145	0	0.501	0.35	0.233	0	0.727
[0.75-0.774]	0.202	0.108	0	0.482	0.403	0.207	0	0.775
[0.775-0.799]	0.178	0.118	0	0.563	0.319	0.222	0	0.671
[0.8-0.824]	0.177	0.106	0	0.418	0.31	0.209	0.043	0.646
[0.825-0.849]	0.125	0.071	0	0.327	0.258	0.184	0	0.589
[0.85-0.874]	0.105	0.131	0	0.418	0.201	0.136	0	0.625
[0.875-0.899]	0.061	0.066	0	0.248	0.179	0.123	0	0.651
[0.9-0.924]	0.039	0.061	0	0.211	0.207	0.156	0	0.614
[0.925-0.949]	0.041	0.067	0	0.248	0.193	0.181	0	0.572
[0.95-0.974]	0.032	0.041	0	0.111	0.099	0.076	0	0.196
[0.975-0.999]	0.005	0.006	0	0.015	0.077	0.152	0	0.519

This table gives the average, standard deviation, minimum and maximum particularity value for the BP comparisons of the case 2. The 20 categories contain all the results that range from a similarity of 0.5 to 0.999 with steps of 0.025.

Table 4. Particularity value statistics in 20 similarity values ranges from case 2 - MF measures

MF	S-value-based particularity				IC-based particularity			
Similarity	Average	Std dev.	Min	Max	Average	Std dev.	Min	Max
[0.5-0.524]	0.341	0.26	0	0.798	0.494	0.162	0.296	0.701
[0.525-0.549]	0.35	0.219	0	0.818	0.429	0.212	0	0.703
[0.55-0.574]	0.364	0.32	0	0.731	0.422	0.265	0	0.849
[0.575-0.599]	0.382	0.265	0	0.694	0.378	0.148	0.125	0.591
[0.6-0.624]	0.242	0.079	0.132	0.47	0.397	0.205	0	0.81
[0.625-0.649]	0.207	0.113	0	0.531	0.302	0.145	0.158	0.475
[0.65-0.674]	0.281	0.106	0.117	0.482	0.609	0.137	0.13	0.806
[0.675-0.699]	0.223	0.181	0	0.562	0.453	0.249	0	0.763
[0.7-0.724]	0.26	0.267	0	0.564	0.389	0.248	0	0.806
[0.725-0.749]	0.179	0.176	0	0.482	0.419	0.211	0	0.763
[0.75-0.774]	0.171	0.177	0	0.371	0.315	0.216	0	0.643
[0.775-0.799]	0.125	0.167	0	0.482	0.33	0.241	0	0.777
[0.8-0.824]	0.063	0.056	0	0.137	0.239	0.218	0	0.574
[0.825-0.849]	0.119	0.13	0	0.415	0.316	0.222	0	0.574
[0.85-0.874]	0.041	0.036	0	0.116	0.266	0.175	0	0.531
[0.875-0.899]	0.045	0.05	0	0.126	0.179	0.093	0.086	0.272
[0.9-0.924]	0.024	0.025	0	0.055	0.163	0.153	0	0.388
[0.925-0.949]	0.02	0.026	0	0.086	0.09	0.107	0	0.272
[0.95-0.974]	0.005	0.007	0	0.023	-	-	-	-
[0.975-0.999]	-	-	-	-	-	-	-	-

This table gives the average, standard deviation, minimum and maximum particularity value for the MF comparisons of the case 2. The 20 categories contain all the results that range from a similarity of 0.5 to 0.999 with steps of 0.025. "-" value denotes an empty category.

Table 5. Particularity value statistics in 20 similarity values ranges from case 2 - CC measures

CC	S-value-based particularity				IC-based particularity			
Similarity	Average	Std dev.	Min	Max	Average	Std dev.	Min	Max
[0.5-0.524]	0.353	0.233	0	0.846	0.621	0.244	0	0.911
[0.525-0.549]	0.36	0.214	0	0.819	0.707	0.15	0.185	0.977
[0.55-0.574]	0.33	0.187	0	0.799	0.64	0.202	0	0.897
[0.575-0.599]	0.341	0.185	0	0.752	0.613	0.194	0	0.896
[0.6-0.624]	0.317	0.183	0	0.754	0.621	0.165	0	0.888
[0.625-0.649]	0.268	0.18	0	0.706	0.592	0.207	0	0.852
[0.65-0.674]	0.28	0.177	0	0.656	0.553	0.227	0	0.888
[0.675-0.699]	0.24	0.177	0	0.583	0.495	0.241	0	0.845
[0.7-0.724]	0.13	0.159	0	0.543	0.466	0.24	0	0.825
[0.725-0.749]	0.196	0.151	0	0.579	0.428	0.268	0	0.82
[0.75-0.774]	0.134	0.122	0	0.484	0.383	0.246	0	0.819
[0.775-0.799]	0.15	0.127	0	0.489	0.391	0.267	0	0.768
[0.8-0.824]	0.144	0.093	0	0.269	0.19	0.187	0	0.625
[0.825-0.849]	0.133	0.123	0	0.421	0.352	0.231	0	0.73
[0.85-0.874]	0.146	0.152	0	0.373	0.255	0.216	0	0.624
[0.875-0.899]	0.051	0.051	0	0.11	0.145	0.152	0	0.381
[0.9-0.924]	0.067	0.085	0	0.269	0.095	0.095	0	0.189
[0.925-0.949]	-	-	-	-	-	-	-	-
[0.95-0.974]	-	-	-	-	0.131	0.131	0	0.262
[0.975-0.999]	0.012	0.012	0	0.024	0.049	0.049	0	0.098

This table gives the average, standard deviation, minimum and maximum particularity value for the CC comparisons of the case 2. The 20 categories contain all the results that range from a similarity of 0.5 to 0.999 with steps of 0.025. "-" value denotes an empty category.

Table 6. Similarity and particularity values of two couples of genes from case 2

SV-based		AQP6	AQP3
Sim	AQP6	1	0.696
	AQP3		1
Par	AQP6	0	0.247
	AQP3	0.415	0

IC-based		AQP6	AQP3
Sim	AQP6	1	0.81
	AQP3		1
Par	AQP6	0	0.531
	AQP3	0.388	0

SV-based		AQP8	AQP5
Sim	AQP8	1	0.704
	AQP5		1
Par	AQP8	0	0
	AQP5	0.19	0

IC-based		AQP8	AQP5
Sim	AQP8	1	0.8
	AQP5		1
Par	AQP8	0	0
	AQP5	0.13	0

The similarity between AQP6 and AQP3 is very close to the similarity between AQP8 and AQP5 regardless the method used (SV or IC-based). However, the particularity profile obtained for each couple is very different. Again, the SV-based and IC-based methods led to the same conclusion.

Table 7. Similarity and particularity pattern in pairwise comparisons on homolog genes in the HomoloGene database

Branch of GO	BP	MF	CC	All
Number of comparisons	1,843,998	1,843,998	1,843,998	5,531,994
Only one gene is annotated	511,899	574,815	581,819	1,668,533
No annotated gene	939,010	823,444	887,419	2,649,873
Two genes annotated	393,089	445,739	374,760	1,213,588
Sim ≥ 0.5 ; All Par < 0.5	287,288	396,412	314,572	998,272
Sim ≥ 0.5 ; One Par ≥ 0.5	39,312	20,754	32,531	92,597
Sim ≥ 0.5 ; Two Spe ≥ 0.5	410	91	54	555
Sim < 0.5	66,079	28,482	27,603	122,164

Green cells refer to valid comparisons where the two genes were annotated.

Table 8. Semantic similarity and particularity values between Exportin-5 orthologs in 9 species

MF			*Homo sapiens* (57510)	*Canis canis* (474913)	*Gallus gallus* (421450)	*Drosophila melanogaster* (32970)	*Macaca mulatta* (700664)	*Mus musculus* (72322)	*Rattus norvegicus* (363194)	*Bos taurus* 100139154)	*Arabidopsis thaliana* (819666)
		Annots	11	4	10	6	10	10	10	10	18
SIM	hsa	11	1	0.428	0.825	0.726	0.825	0.885	0.797	0.825	0.209
	hsa	4		1	0.676	0.693	0.676	0.516	0.585	0.676	0.043
	cca	10			1	0.581	1	0.957	0.862	1	0.06
	gga	6				1	0.581	0.509	0.591	0.581	0.214
	mmul	10					1	0.957	0.862	1	0.06
	mmu	10						1	0.849	0.957	0.072
	rno	10							1	0.862	0.234
	bta	10								1	0.06
	ath	18									1
PAR	hsa	11	0	0.611	0.082	0.308	0.082	0.082	0.082	0.082	0.641
	cca	4	0	0	0	0	0	0	0	0	0.661
	gga	10	0	0.52	0	0.336	0	0	0	0	0.707
	dme	6	0	0.328	0.119	0	0.119	0.119	0.119	0.119	0.444
	mmul	10	0	0.52	0	0.336	0	0	0	0	0.707
	mmu	10	0	0.52	0	0.336	0	0	0	0	0.707
	rno	10	0	0.52	0	0.336	0	0	0	0	0.707
	bta	10	0	0.52	0	0.336	0	0	0	0	0.707
	ath	18	0.871	0.947	0.905	0.871	0.905	0.905	0.905	0.905	0

Color gradient according to similarity value (0 = white, 1 = blue) and particularity values (0 = white, 1 = red or green). If Par(gene1, gene2) is displayed in green, Par(gene2, gene1) is displayed in red. The value contained in a cell is the particularity of the gene displayed at its row header compared to the gene displayed at its column header. The given numbers of annotations (#Annot) consider the total number of GO terms that annotate the genes either directly or indirectly).

3 Synthèse

Nous avons proposé une mesure de particularité sémantique. Cette mesure repose sur la notion d'informativité, qui est compatible avec les approches basées sur le contenu d'information aussi bien qu'avec la valeur sémantique de l'approche de Wang. Nous avons démontré l'utilité de la mesure de particularité sémantique, notamment pour identifier et quantifier des caractéristiques propres à un produit de gène comparé à des produits de gènes similaires.

Cette mesure ne remplace pas une mesure de similarité, mais devrait être utilisée conjointement à une telle mesure. En effet, la mesure de similarité sémantique est symétrique. Ce n'est pas le cas de la mesure de particularité sémantique, puisque la particularité mesurée en comparant A à B est généralement différente de la particularité réciproque. Lorsque l'on compare deux gènes, on obtient donc des triplets (similarité, particularité, particularité réciproque).

Les résultats de comparaison qui nous intéressent le plus sont ceux présentant une forte valeur similarité et une forte valeur de particularité parmi les deux obtenues. Cependant, ces cas ne sont pas détectables en utilisant seulement une mesure de similarité sémantique. Or dans le cadre d'une comparaison inter-espèces, ils permettent d'identifer des fonctions propres à une espèce au sein d'un métabolisme qui paraît au premier abord simplement « similaire ». La comparaison sémantique de produits de gènes repose donc sur l'interprétation des triplets obtenus en utilisant une mesure de similarité et notre mesure de particularité.

CHAPITRE 4

INTERPRÉTATION DES RÉSULTATS D'UNE MESURE SÉMANTIQUE

Dans ce chapitre, nous présentons une méthode permettant de définir un seuil de similarité et un seuil de particularité. Ces seuils permettent de savoir à partir de quelle valeur de similarité il est possible de considérer deux gènes comme similaires, et à partir de quelle valeur de particularité il est possible de considérer que les particularités d'un gène sont significatives. Ces seuils nous permettront d'interpréter les triplets de similarité / particularités définis dans le chapitre précédent. En effet, grâce à ces seuils, il nous sera possible de distinguer systématiquement les gènes ayant des fonctions particulières parmi des gènes similaires, qui sont des cas particulièrement intéressants dans le cadre de la comparaison inter-espèces de voies métaboliques.

Sommaire

1 Introduction

L'interprétation qualitative des triplets de valeurs similarité et particularités est difficile. En effet, deux questions se posent. D'une part, à partir de quelle valeur de similarité deux produits de gènes peuvent être considérés comme similaires. D'autre part, quelle valeur de particularité marque l'existence d'une fonction différente. Accepter ou non la transposition de résultats biologiques entre espèces demande d'y répondre. En effet, si les gènes intervenant dans une voie métabolique sont à la fois similaires entre deux espèces et présentent peu de particularités, la transposition est cohérente. L'article suivant en cours d'examen à PLoS ONE répond à cette problématique.

2 Article

Thresholds of semantic similarity and particularity for gene set functional analysis

- Résumé -

L'analyse des termes GO qui annotent les gènes joue un rôle important dans l'interprétation des données issues de processus à haut-débits. Cette analyse met typiquement en œuvre des mesures de similarité et particularité sémantiques capables de quantifier l'importance des annotations GO. Cependant, il n'existait pas jusqu'à ce jour de méthode capable de valider l'interprétation de valeurs de similarité et de particularité de façon à déterminer si deux gènes ou ensembles de gènes sont similaires ou si un gène possède une fonction particulière significative. Cette interprétation est souvent basée soit sur un seuil implicite, soit sur un seuil arbitraire (typiquement : 0.5). Cet article présente une méthode pour déterminer des seuils de similarité et de particularité. Nous avons comparé des distributions de valeurs de similarités issues de la comparaison de gènes que l'on savait similaires et de gènes que l'on savait non similaires. Nous avons procédé à cette comparaison sur les trois branches de Gene ontology. Dans toutes les situations, nous avons observé un chevauchement entre les distributions similaires et non similaires, indiquant que des gènes similaires pouvaient avoir une valeur de similarité plus basse que des gènes non similaires. Nous avons proposé une méthode pour déterminer les seuils optimaux de similarité et particularité en minimisant respectivement les proportions de faux positifs et faux négatifs dans les valeurs de similarités, et les proportions de triplets (similarité, particularité, particularité réciproque) peu informatifs. Nous avons évalué nos seuils sur la totalité de la base de données HomoloGene. Pour chaque groupe de gènes homologues, nous avons calculé toutes les valeurs de similarités et de particularités entre les gènes pris deux par deux. Enfin, nous avons ciblé la famille multigénique PPAR et nous avons montré que les triplets de résultats de nos mesures formaient des motifs permettant de mieux discriminer les orthologues des paralogues. Nous proposons une méthode pour déterminer les seuils optimaux de similarité et de particularité pour Gene Ontology. Leur utilisation résulte en la formation de différents motifs de similarité et particularité. L'analyse qualitative menée sur la famille multigénique PPAR a montré que ces seuils permettent d'obtenir des motifs biologiquement pertinents.

Thresholds of semantic similarity and particularity for gene set functional analysis

Charles Bettembourg[1,2,3,4,5*], Christian Diot[2,3], Olivier Dameron[1,4,5]
1 Université de Rennes 1, 35000 Rennes, France
2 INRA, UMR1348 PEGASE, Saint-Gilles, France
3 Agrocampus OUEST, UMR1348 PEGASE, Rennes, France
4 IRISA, Campus de Beaulieu, 35042 Rennes, France
5 INRIA, France
∗ **E-mail: charles.bettembourg@univ-rennes1.fr**

Abstract

Background: The analysis of genes' Gene Ontology annotations plays an important role in the interpretation of high throughput experiments results. This analysis typically involves semantic similarity and particularity measures that quantify the importance of the Gene Ontology annotations. However, there is currently no sound method supporting the interpretation of the similarity and particularity values in order to determine whether two sets of genes are similar or whether one gene has some significant particular function. This interpretation is frequently based either on an implicit threshold, or an arbitrary one (typically 0.5). This article focuses on a method for determining the similarity and particularity thresholds.

Results: We compared the distributions of the similarity values of pairs of similar genes and of pairs of non-similar genes. We performed these comparisons separately for the three branches of the Gene Ontology. In all the situations, we observed an overlap between the similar and the non-similar distributions, indicating that some similar genes had a similarity value lower than the similarity value of some non-similar genes. We proposed a method for determining the optimal similarity and particularity thresholds by minimizing the proportions of similarity false positives and of false negatives and by minimizing the proportions of undesirable patterns, respectively. We evaluated our thresholds on the whole HomoloGene database. For each group of homologue genes, we computed all the similarity and particularity values between pairs of genes. Finally, we focused on the PPAR multigenic family and showed that the similarity and particularity patterns obtained with our thresholds were better at discriminating orthologs and paralogs.

Conclusion: We proposed a method for determining optimal semantic similarity and particularity thresholds on the Gene Ontology. Using them results in different similarity and particularity patterns. The qualitative analysis on the PPAR multigenic family showed that these threshold yielded biologically-relevant patterns.

Introduction

Comparing several gene sets and identifying and quantifying their common features as well as the ones that differentiate them are important parts of gene sets functional analysis [1–3]. These operations hinge on the comparison of sets of Gene Ontology (GO) terms [4]. Numerous semantic similarity measures have been developed [5–7]. Recently, we have proposed to combine semantic similarity measures and a new semantic particularity measure to improve the results of gene sets analysis. The analysis of the similarity and particularity results is based on an interpretation that contrasts the genes having particular functions among similar genes. Previous studies have mainly focused on the definitions of measures. However, there is no extensive study about the interpretation of theses values. As a result, interpretation is frequently based either on an implicit threshold or an arbitrary one. Moreover, the value of these threshold may vary over time, as both GO and GOA evolve. In this study, we analyze how similarity and particularity

values are distributed and we propose adequate thresholds.

The GO terms annotating genes describe the biological processes, molecular functions and cellular components each gene is involved in. If these terms were independent, gene set characterization could be performed by a straightforward set-based approach such as the Jaccard index or Dice's coefficient. However, GO terms are hierarchically-linked. Consequently, the characterization needs to take into account the underlying ontological structure of the GO annotations [8]. Several semantic similarity measures that exploit the formal representation of the meaning of the terms by considering the relations between the terms have been developed and evaluated [5]. Pesquita et al. classified these measures in two categories: node and edge-based methods, with some hybrid measures. Node-based measures assign each ontology term an Information Content (IC) value, the least frequent terms having the highest IC value. This IC concept, borrowed from Shannon's Information Theory [9], was used to measure similarities using ontologies [10–12] such as WordNet [13]. Node-based measures consider that the similarity between two terms rely on their most informative common ancestor. These methods developed in linguistics have been applied to GO [14, 15], the IC of a GO term being inversely proportional to the frequency with which it annotates a gene using the Gene Ontology Annotations (GOA) database [16]. In the context of genes comparison, IC-based methods have three main limits related to their dependence on a GOA-based corpus. First, it can prove difficult or even impossible to obtain a relevant corpus. GOA provides single and multi-species tables of annotation. Although using a species-specific table is well-suited to intra-species comparisons, it becomes problematic for inter-species comparisons. Second, using a multi-species table (like the UniprotKB table) for cross-species studies is biased towards the most extensively annotated species such as human or mice. Third, the well-studied areas of biology have high annotation frequencies and are therefore less informative and see their importance downgraded, whereas the less-studied areas are artificially upgraded [17–19]. Edge-based measures compute a distance between GO terms using the directed graph topology. This distance can be the shortest path between two compared terms [20] or the length of the path between the root of the ontology and the lowest common ancestor of the compared terms [21–25]. This last distance makes terms with a deep common ancestor more similar than terms with a common ancestor close to the root. Unlike node-based measures, edge-based measures are not corpus-dependent. However, granularity is not uniform in GO, so terms at the same depth can have different precisions [26]. Hybrid methods combine different aspects of node-based and edge-based methods. Wang *et al.* method assigns each term a semantic value that represents how informative the term is, conforming to the node-based approach [27]. However, the semantic value of a term is obtained by following the path from this term to the root and summing the semantic contributions of all the ancestors of this term. As the semantic value depends on the ontology topology, it also conforms to the edge-based approach. Most of these methods are designed to compare terms but not sets of terms (as needed to compare genes). Common approaches proposed to compare genes consider the average [14], the maximum [28] of all pairwise similarities, or only the best matching pairs [29, 30]. Pesquita et al. consider that the best-match average variants are the best overall. However, they do not single out any specific semantic similarity measure as the best one, because the optimal measure will depend on the data to compare and the level of detail expected in the results. The main advantage of Wang's method compared to purely node-based methods is that unlike the IC, the semantic value is not GOA-dependent. It is thus well-suited to cross-species comparisons. Semantic similarity measures typically focus on what is common between the two compared entities. We recently developed a semantic particularity measure to also take into account what distinguishes each compared entity from the other one [31]. The semantic particularity of a set A of GO terms compared to another set B of GO terms depends on the informativeness measure of the A terms that are not in B. This informativeness measure can be Wang's semantic values or an Information Content value. This concept of particularity is to use in combination with a semantic similarity in order to improve gene set functional analysis.

The data analysis often hinges on a qualitative interpretation of the similarity values in order to contrast similar and dissimilar pairs of genes. This discretization of the similarity and particularity

values makes the interpretation easier. It helps to consider whether a functional difference is marginal or not while comparing two genes. However, no systematic analysis of the optimal threshold value separating similar from dissimilar has been conducted. Some studies avoid the problem by focusing only on high or low values (without mentioning when a value reaches this point). Other studies draw the line at 0.5 (with no other motivation than 0.5 being the mid-range value of the similarity interval).

The main factors influencing the similarity values are: granularity differences in GO, GO topological differences between BP, MF and CC, quantity and quality of gene annotations, GO temporal evolution [32]. In some cases, the threshold of 0.5 may be unadapted. For example, the similarity value between the protein tyrosine kinase 2 (PTK2) and the Ubiquitin B (UBB) is 0.502 using Wang similarity measure on their Biological Processes (BP) annotations. This value is just above the mid-interval intuitive threshold. These two genes are well annotated: they have respectively 73 and 79 distinct BP annotations. According to Entrez Gene, PTK2 is involved in cell growth and intracellular signal transduction pathways triggered in response to certain neural peptides or to cell interactions with the extracellular matrix while UBB is required for ATP-dependent, nonlysosomal intracellular protein degradation of abnormal proteins and normal proteins with a rapid turnover. These processes can not be considered as similar. Consequently, the 0.502 value of similarity does not allow to consider PTK2 and UBB as similar genes according to the BP they participate in. There is a need for a systematic study of the semantic measures values in order to determine optimal similarity and particularity thresholds for the qualitative part of gene set functional analysis.

Method

We propose a method to define a threshold for a node-based and a hybrid semantic similarity measure as well as corresponding semantic particularity measures.

Metrics

Semantic similarity

Lin's Method. Lin is a widely used node-based similarity measure method that uses the Information Content (IC) concept [11]. Several tools available have implemented this method. The IC of a term t depends on its log probability $P(t)$. Working with the Gene Ontology terms, this IC is inversely proportional to the frequency with which the terms annotates a gene using the Gene Ontology Annotations (GOA) database. While comparing two GO terms t1 and t2, having a most informative common ancestor t0, Lin defines their similarity as follows:

$$Sim(t1, t2) = \frac{2 \times logP(t0)}{logP(t1) + logP(t2)}$$

In this article, we computed Lin's similarity with the GOSemSim R package using the best-match average approach to compare genes [33].

Wang's method. This method depends only on the GO graph and does not need any annotation corpus, allowing cross-species comparisons [27]. The first step of the method is to compute the semantic contributions of the ancestors of each term to compare to these terms, following:

$$\begin{cases} S_A(A) = 1 \\ S_A(t) = max\{w_e * S_A(t') \mid t' \in children\ of\ (t)\}\ if\ t \neq A \end{cases}$$

where $S_A(t)$ is the semantic contribution of the term t to the term A and w_e is the semantic contribution factor for edge e linking a term t with its child term t'. According to Wang, we use a semantic

contribution factor of 0.8 for the "is a" relations and 0.6 for the "part of" relations, and we added a 0.7 factor for the "[positively] [negatively] regulates" relations. Then, for each target term to compare, the semantic value is the sum of the semantic contributions of all its ancestors:

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

The comparison of two terms A and B is computed as following:

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

The similarity between a term "go" and a set of term "GO" is:

$$Sim(go, GO) = \max_{1 \leqslant i \leqslant k} (S_{GO}(go, go_i))$$

Last, the similarity between two genes G1 and G2 is:

$$Sim(G1,G2) = \frac{\sum_{1 \leqslant i \leqslant m} (Sim(go_{1i}, GO_2) + \sum_{1 \leqslant j \leqslant n} (Sim(go_{2j}, GO_1)}{m+n}$$

In this article, we used an in-house implementation of Wang's similarity computation.

Semantic particularity

In a previous article, we defined the semantic particularity of a set of GO terms Sg1 compared to another set of GO terms Sg2 [31].

Some of the terms of Sg1 that are not members of Sg2 may be linked in the graph. Taking several linked terms into account would result in considering them several times. Therefore, the particularity measure only focuses on the terms of Sg1 that do not have any descendant in Sg1 and that are not members of Sg2. Some of these terms might be ancestors of terms of Sg2 and should be considered as common to Sg1 and Sg2. Sg^* is the union of Sg and the sets of ancestors of each element of Sg. MPT(Sg1, Sg2) is the set of most particular terms of Sg1 compared to Sg2, i.e. the set of terms of Sg1 that do not have any descendant in Sg1 and that are not members of $Sg2^*$. PI(Sg1, Sg2) is the particular informativeness of a set of GO terms Sg1 compared to another set of GO terms Sg2, i.e. the sum of the differences between the informativeness (I) of each term t_p of MPT(Sg1, Sg2) and the informativeness of the most informative common ancestor (MICA) between t_p and Sg2. The PI of a set of terms is the information that is not shared with the other set.

$$PI(Sg1, Sg2) = \sum_{t_p \in MPT(Sg1,Sg2)} I(t_p) - I(MICA(t_p, Sg2))$$

PI is normalized to compute Par(Sg1, Sg2), the semantic particularity of the set of GO terms Sg1 compared to the set of GO terms Sg2. MCT(Sg1, Sg2) is the set of the most informative common terms of Sg1 and Sg2, i.e. the set of the terms belonging to the intersection of $Sg1^*$ and $Sg2^*$ that do not have any descendant either in $Sg1^*$ or in $Sg2^*$. Par(Sg1, Sg2) is the ratio of PI(Sg1, Sg2) and the sum of the informativeness of Sg1 most informative terms (i.e. those Sg1-specific and those common with Sg2; the MICA in the PI formula for the Sg1-specific guarantees that the informativeness of common terms is not counted twice).

$$Par(Sg1, Sg2) = \frac{PI(Sg1, Sg2)}{PI(Sg1, Sg2) + \sum_{t_c \in MCT(Sg1,Sg2)} I(t_c)}$$

Similarity threshold determination

Interval determination for similarity threshold

Ideally, the similarity threshold would allow to distinguish all similar and non-similar genes, without false positives nor false negatives. When comparing two similarity distributions, one of similar genes (S) and one of non-similar genes (N), the minimum value of S should be greater than the maximum value of N. S and N distributions can be represented by boxplots, which may overlap.

Figure 1 illustrates the case without overlap, where we have min(S) = a and max(N) = b. On this Figure, $a > b$. A similarity value greater than a means that the compared genes are similar. A similarity value lower than b means that the compared genes are not similar. A similarity value between a and b means that the compared genes are nearly similar and an expert opinion could be required to specify the result.

Figure 2 illustrates the case where the S and N distributions overlap, meaning that there are some false positives (i.e. pairs of genes from N that are not similar but that have a similarity value greater than a) and false negatives (i.e. pairs of genes from S that are similar but have a similarity value lower than b). In this case, a similarity value lower than a means that the compared genes are not similar. A similarity value greater than b means that the compared genes are similar. Again, an expert opinion could be required to specify the result in this interval. However, it is possible in this case to find the threshold value that minimizes both FP and FN.

It is possible to establish a general decision framework which will work in the two cases described in this section. We can define three thresholds values:

- $\tau_S = \max(a, b)$ is the threshold value above which the two compared genes are similar. There can not be any FP above τ_S, but there may be some FN below τ_S if $a < b$.
- $\tau_N = \min(a, b)$ is the threshold value under which the two compared genes are not similar. There cannot be any FN below τ_N, but there may be some FP above τ_N if $a < b$.
- τ_{sim} is the threshold value located between τ_S and τ_N that that minimizes the proportion of FP and FN. The closer τ_{sim} is to τ_S, the more FN and the fewer FP (and conversely).

BP, MF and CC similarity thresholds

In order to determine τ_S, τ_N and τ_{sim}, we constituted different S and N distributions for BP, CC and MF.

For BP, we computed similarity values between all the pairs of genes from a same PANTHER family to obtain an S distribution. The PANTHER (Protein ANalysis THrough Evolutionary Relationships) database classifies proteins (and their genes) in order to facilitate high-throughput analysis [34]. The PANTHER families are composed of genes sharing an evolution history, molecular functions and biological processes annotations, and an involvment in the same biological pathways. We assumed that genes belonging to a same PANTHER family share enough features to be considered as being involved in similar biological processes. We computed six S distributions from six different PANTHER families: histone h1/h5 (pthr11467), g-protein coupled receptor (pthr12011), neurotransmitter gated ion channel (pthr18945), tyrosine-protein kinase receptor (pthr24416), phosphatidylinositol kinase (pthr10048) and sulfate transporter (pthr11814). We computed fifteen N distributions corresponding to all the combinations of two families among the previous six. Each N distribution is composed of the similarity values

between a gene from the first family and a gene from the second one. We assumed that two genes belonging to two different PANTHER families should not be considered as being involved in similar biological processes.

For MF, we used the same six genes families to compute our six S and our fifteen N distributions, as the PANTHER families are also homogeneous in term of molecular functions.

For CC, we used the genes from five different pathways, each one located in a different cellular compartment, to compute our five S and ten N distributions. The lists of genes were provided by the Reactome database [35]. The five pathways and their respective compartment were were: chromosome maintenance (nucleoplasm and nuclear membrane), mitochondrial protein import (mitochondrial inter-membrane space, membrane and matrix), potassium channel (cellular membrane), protein folding (cytosol), termination of O-glycan biosynthesis (Golgi lumen).

Particularity threshold

The result of a semantic comparison of genes is a tuple of one similarity value and two particularity values. Combining a similarity measure with a particularity measure allows to classify the results of our comparisons using the eight distinct patterns described in table 1. A comparison should not result in a "+ + +" or a "- - -" pattern. Indeed, a "+ + +" pattern would mean that the two compared genes share enough features to be considered as similar and, in the same time, they have respectively enough particular features to be both considered as particular. Conversely, a "- - -" pattern would mean that the two compared genes are neither similar, nor particular. Consequently, we can refine the different threshold intervals we proposed before by minimizing the "+ + +" and "- - -" results while varying the thresholds inside these intervals.

Threshold stability study

We validated our study using a leave-one-out approach which consisted in successively recomputing the thresholds using all the sets but one. This approach allows to assess the stablility of the thresholds.

Evaluation

The evaluation study consisted in both quantifying the extent of the changes resulting from using the threshold we computed instead of the default 0.5, and in determining whether these changes are biologically-relevant.

The first part of this study focused on the changes in the results of the whole HomoloGene database intra-group genes comparisons.

In the second part of this study, we computed the similarity and particulartity measures on the well annotated PPAR multigenic family (PPARα, PPARβ and PPARγ). Our goal was to determine whether our similarity and particularity thresholds leads to biologically more relevant interpretations than the default approach. In the PPAR family, we considered the distinction between the orthologs and the paralogs.

Results and discussion

Determination of a threshold range using semantic comparisons of genes

We studied the similarity and particularity values obtained while comparing genes known to be functionally close and genes without functional proximity. We used two different semantic similarity measures: an hybrid (Wang) and a node-based (Lin). We used the semantic particularity measure of Bettembourg et al., respectively with Semantic value and with IC.

Figure 3 presents the distribution of the BP similarity values obtained for six intra-PANTHER families comparisons and the corresponding fifteen inter-families comparisons. As expected, similarity values obtained using either the Wang (Part A) or Lin (Part B) method in the intra-families comparisons were significantly higher than the ones of the inter-families comparisons , as determined by the Welch test (Supplementary file 1). In order to consider all the intra-families results as similar, the similarity threshold should be lower than the lowest whisker of the intra-families blue box (τ_S): respectively 0.164 for Wang and 0.325 for Lin. In order to consider all the inter-families results as non similar, the similarity threshold should be greater than the upmost whisker of the inter-family yellow box (τ_N): respectively 0.618 for Wang and 0.794 for Lin. Intra-families values lower than the threshold should be considered as False Negatives (FN) and inter-families values greater than the threshold should be considered as False Positives (FP).

Figure 4 presents the distribution of the MF similarity values obtained for our six intra-PANTHER families comparisons and the corresponding fifteen inter-families comparisons. Again and as expected, similarity values obtained using Wang (Part A) or Lin (Part B) method in the intra-groups similarity were significantly higher than the inter-groups ones, as determined by the Welch test (Supplementary file 2). In order to consider all the intra-pathways results as similar, the similarity threshold should be lower than the lowest whisker of the intra-pathways blue box (τ_S): respectively 0.251 for Wang and 0.506 for Lin. In order to consider all the inter-pathways results as non similar, the similarity threshold should be greater than the upmost whisker of the inter-pathways yellow box (τ_N): respectively 0.671 for Wang and 0.725 for Lin.

Figure 5 presents the distribution of the CC similarity values obtained for our five intra-pathways comparisons and the corresponding ten inter-pathways comparisons. Similarity values obtained using either the Wang (Part A) or Lin (Part B) method in the intra-groups similarity were again significantly higher than the inter-groups ones as determined by the Welch test (Supplementary file 3). In order to consider all the intra-pathways results as similar, the similarity threshold should be lower than the lowest whisker of the intra-pathways blue box (τ_S): respectively 0.166 for Wang and 0.28 for Lin. In order to consider all the inter-pathways results as non similar, the similarity threshold should be greater than the upmost whisker of the inter-pathways yellow box (τ_N): respectively 0.773 for Wang and 0.938 for Lin.

For BP, MF and CC, the values obtained with the yellow box were greater than the values obtained with the blue box for both Wang and Lin measures. Consequently, we have in each case an overlap between our S and N distributions, which corresponds to the situation shown in Figure 2.

Threshold value optimization

Overlap study for the determination of a similarity threshold

In the previous section, we have defined the τ_S and τ_N values from Figure 2 for our BP, MF and CC S and N sets. A similarity value lower than τ_N means that the compared genes are not similar whereas a similarity measure result above τ_S means that the compared genes are similar. Between these two values, it is complicated to determine whether a similarity value indicates that the compared genes are similar or not. As we have an overlap between our blue and yellow boxes, defining a threshold in this interval will yield some False Positive and some False Negative results. We have determined the optimal similarity threshold value that minimizes the sum of FP and FN proportions. We used the proportions because S and N distributions have different sizes. Figure 6 displays the results for Wang's SV-based measure and Figure 7 for Lin's IC-based measure. The minimum ordinate value of each curve of the Figures 6 and 7 gives the threshold for BP, MF and CC using respectively the Wang's method and the Lin's method. All the values obtained for the boxplots (Figures 3, 4 and 5) and the threshold variation curves (Figures 6 and 7) are summarized in the Table 2. These similarity thresholds are different depending on the similarity measure used. They also differ between BP, MF and CC. This can be explained by the different complexity level between these three branches [32]. According to the accuracy needed when

interpreting semantic similarity results, it is possible to use one of the three proposed thresholds (τ_N, τ_S and τ_{sim}). None of these thresholds is equal to the intuitive threshold of 0.5.

Particularity threshold

We studied the variation of + + + and - - - profiles in our datasets using one of our three similarity thresholds and varying the value of τ_{par}, the particularity threshold. Let sim be the result of a semantic similarity measure between two genes G1 and G2. If sim is lower than τ_N, we can conclude that G1 and G2 are strictly non-similar. Conversely, if sim is greater than τ_N, we can only conclude that G1 and G2 are possibly similar but with no certainty. If sim is greater than τ_S, we can conclude that G1 and G2 are strictly similar. Conversely, if sim is lower than τ_S, we only can conclude that these genes are possibly non-similar but with no certainty. Last, using τ_{sim} cannot lead to a conclusion with absolute certainty, but has the smallest risk of error. Indeed, using τ_N and τ_S can resut respectively in a great amout of FP and in a great amount of FN. Consequently, we used the similarity threshold τ_{sim} to compute τ_{par}, the particularity threshold.

The Figure 8 displays the results using a SV-based approach and the Figure 9 using an IC-based one. Just like for similarity, we normalized all the + + + and - - - values in percentage of each considered dataset. Table 3 give the particularity thresholds (τ_{par}) minimizing the sum of + + + and - - - patterns.

These thresholds are different between BP, MF and CC and between the different approaches.

Thresholds robustness

In order to study the robustness of our optimization, we successively removed one gene set from our data sets and re-computed the similarity and particularity thresholds. We performed this analysis on BP, MF and CC. Tables 4 and 5 presents the resultsrespectively for SV-based and IC-based methods. According to the different datasets, the thresholds varied slightly.

BP similarity was between 0.4 and 0.435 and BP particularity threshold was between 0.49 and 0.515.

MF similarity threshold stayed stable at 0.41, except when not taking into account the family of the genes relative to the neurotransmitter gated ion channels (0.49). The case of MF was different from BP and CC regarding its similarity (FP + FN) curve. Indeed, the 0.41 minimum value was located at the extreme left of a part of the curve where (FP + FN) varied very slightly. Consequently, leaving out the dataset "neurotransmitter gated ion channels" which was causing this specific position of minimum greatly affected the threshold. However, it should be put in perspective first with the relatively long interval in which the sum of FP and FN remained low, and second, with the fact that the minimum of 0.49 obtained without the "neurotransmitter gated ion channels" set was located at the opposite part of this range of stability. MF particularity threshold was between 0.35 and 0.485.

CC similarity threshold was between 0.475 and 0.515 and CC particularity threshold was between 0.28 and 0.335.

Considering Figures 6, 7, 8 and 9, the minimum ordinate value of the sums FP + FN and "+ + +" + "- - -" was located in each case in a relatively large range in which the ordinate varied only slightly. Consequently, we concluded that the similarity and particularity thresholds can be located in the range where the sum of the FP, FN, + + + and - - - proportions varied the least. Last, we have to notice that each threshold presented here admits errors in the proportion described in the Table 4.

Evaluation

Large scale evaluation of the thresholds changes impact

We evaluated the impact of our new similarity and particularity thresholds over a large dataset characterization. We compared the repartition of semantic measures results among the different patterns proposed in table 1 for the whole HomoloGene database considering a 0.5 arbitrary threshold and our thresholds.

HomoloGene is a system that automatically detects homologs, including paralogs and orthologs, among the genes of 21 completely sequenced eukaryotic genomes [36]. The tables 6, 7 and 8 summarize the results respectively for BP, MF and CC. We have not distinguished the "+ + -" and "+ - +" categories as well as the "- + -" and "- - +", because the order of particularities values in the results of this study is meaningless. All category of the pattern described in Table 1 have been impacted by the change of threshold. As the new thresholds are different between BP, MF and CC, the transitions observed are different. For example, the number of "+ + -" increased for BP while it decreased for MF and CC. However, in all cases, the greatest increase of effective concerns the "+ + - and + - +" category, respectively +26.2%, +18.5% and + 36.7% for BP, MF and CC. The number of "+ + +" and "- - -" cases, that are the less informative ones, decreased for BP (-11.2%) and MF (-34.8%) but increased for CC (+49%). Overall, the change of thresholds deeply impacted the repartition of the HomoloGene intra-groups comparison results between the different patterns.

Case study: the PPAR multigenic family

The multigenic family of the peroxisome proliferator activated receptor (PPAR) is involved in different processes [37] as a transcription factor. Each member of this family (PPARα, PPARβ and PPARγ) uses the same molecular mechanisms in different metabolisms. This family was well conserved through evolution [38]. We expected a similarity value above the threshold for BP when comparing PPAR orthologs in several species. However, the ortholog conjecture assume that orthologs generally share more functions than paralogs. We consequently expected some similarity values under the threshold when comparing PPAR paralogs inside a species and between species. Table 8 and 9 displays the results of this study respectively for BP and MF. Each gene was only annotated by one or two CC terms, so we kept CC results out of this study. All our similarity values were greater than τ_{sim}. To observe some differences between orthologs and paralogs similarity, we consequently had to use τ_S. This threshold gives the certainty that the results above it indicate two similar genes. However, the only conclusion that can be inferred for the genes comparisons resulting in values between τ_{sim} and τ_S is that there is a doubt on these genes being similar. The results of inter-orthologs comparisons systematically matched a "+ - -" pattern, according to our expectations. In contrast, we observed some values lower than τ_S and greater than τ_{par} when comparing paralogs, resulting in "+ + -", "- + -" and "- - +" patterns. A recent article of Thomas *et al.* "strongly encourage careful consideration of the interpretations" of GO-related analysis [39]. Consequently, the only possible conclusion here is that the actual state of the PPAR annotation is consistent with the ortholog conjecture, according to a similarity and a particularity measure, using our new thresholds.

Conclusion

In this article, we proposed a method for determining the similarity and particularity thresholds for BP, MF and CC branches of Gene Ontology. These new thresholds allow a new insight over semantic measure results. We showed that the results of the comparisons in the HomoloGene database were classified very differently using the new thresholds. These new thresholds also better separated orthologs and paralogs in the multigenic family of PPAR. The new thresholds we proposed are not absolute. As the curve used to define them draw a plane where the thresholds are minimized, we can pick our thresholds in a relatively large range. The precise thresholds values we proposed are only the minimum value of this range. Furthermore, we think that a threshold value as to be considered in a biological context. It has to be reevaluated considering this context, GO and GOA evolution and the semantic measure used.

Acknowledgments

CB was supported by a fellowship from the French ministry of research.

References

1. Grossmann S, Bauer S, Robinson PN, Vingron M (2007) Improved detection of overrepresentation of gene-ontology annotations with parent child analysis. Bioinformatics 23: 3024–31.

2. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1–13.

3. Barriot R, Sherman DJ, Dutour I (2007) How to decide which are the most pertinent overly-represented features during gene set enrichment analysis. BMC Bioinformatics 8: 332.

4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. Nat Genet 25: 25–9.

5. Pesquita C, Faria D, Falcão AO, Lord P, Couto FM (2009) Semantic similarity in biomedical ontologies. PLoS Comput Biol 5: e1000443.

6. Gan M, Dou X, Jiang R (2013) From ontology to semantic similarity: calculation of ontology-based semantic similarity. ScientificWorldJournal 2013: 793091.

7. Wu X, Pang E, Lin K, Pei ZM (2013) Improving the measurement of semantic similarity between gene ontology terms and gene products: insights from an edge- and ic-based hybrid method. PLoS One 8: e66745.

8. Rhee SY, Wood V, Dolinski K, Draghici S (2008) Use and misuse of the gene ontology annotations. Nat Rev Genet 9: 509–15.

9. Shannon CE (1948) A mathematical theory of communication. Bell system technical journal 27.

10. Resnik P (1999) Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence 11: 95–130.

11. Lin D (1998) An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning : 296–304.

12. Jiang J, Conrath D (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the International Conference Research on Computational Linguistics (ROCLING). Taiwan.

13. Fellbaum C (1998) WordNet: An Electronic Lexical Database. MIT Press.

14. Lord PW, Stevens RD, Brass A, Goble CA (2003) Semantic similarity measures as tools for exploring the gene ontology. In: Pacific Symposium on Biocomputing. pp. 601–612.

15. Sheehan B, Quigley A, Gaudin B, Dobson S (2008) A relation based measure of semantic similarity for gene ontology annotations. BMC Bioinformatics 9: 468.

16. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, et al. (2004) The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. Nucleic Acids Res 32: D262–6.

17. Jin B, Lu X (2010) Identifying informative subsets of the gene ontology with information bottleneck methods. Bioinformatics 26: 2445–51.

18. Gillis J, Pavlidis P (2013) Assessing identity, redundancy and confounds in gene ontology annotations over time. Bioinformatics 29: 476–82.

19. Chen G, Li J, Wang J (2013) Evaluation of gene ontology semantic similarities on protein interaction datasets. Int J Bioinform Res Appl 9: 173–83.

20. Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man, and Cybernetics 19: 17–30.

21. Pekar V, Staab S (2002) Taxonomy learning - factoring the structure of a taxonomy into a semantic classification decision. In: COLING.

22. Wu Z, Palmer M (1994) Verb semantics and lexical selection. In: Proc. of the 32nd annual meeting on Association for Computational Linguistics. pp. 133–138. doi: http://dx.doi.org/10.3115/981732.981751.

23. Cheng J, Cline M, Martin J, Finkelstein D, Awad T, et al. (2004) A knowledge-based clustering algorithm driven by gene ontology. J Biopharm Stat 14: 687–700.

24. Alvarez MA, Yan C (2011) A graph-based semantic similarity measure for the gene ontology. J Bioinform Comput Biol 9: 681–95.

25. Díaz-Díaz N, Aguilar-Ruiz JS (2011) Go-based functional dissimilarity of gene sets. BMC Bioinformatics 12: 360.

26. Mazandu GK, Mulder NJ (2012) A topology-based metric for measuring term similarity in the gene ontology. Adv Bioinformatics 2012: 975783.

27. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF (2007) A new method to measure the semantic similarity of go terms. Bioinformatics 23: 1274–81.

28. Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, et al. (2005) Correlation between gene expression and go semantic similarity. IEEE/ACM Trans Comput Biol Bioinform 2: 330–8.

29. Couto FM, Silva MJ, Coutinho P (2005) Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. In: Herzog O, Schek HJ, Fuhr N, Chowdhury A, Teiken W, editors, CIKM. ACM, pp. 343–344. URL `http://dblp.uni-trier.de/db/conf/cikm/cikm2005.html#CoutoSC05`.

30. Azuaje F, Wang H, Zheng H, Bodenreider O, Chesneau A (2006) Predictive integration of gene ontology-driven similarity and functional interactions.

31. Bettembourg C, Diot C, Dameron O (2013) Semantic particularity measure for functional characterization of gene sets using gene ontology. PLoS One (In review) .

32. Dameron O, Bettembourg C, Le Meur N (2013) Measuring the evolution of ontology complexity: the gene ontology case study. PLoS One 8: e75993.

33. Yu G, Li F, Qin Y, Bo X, Wu Y, et al. (2010) Gosemsim: an r package for measuring semantic similarity among go terms and gene products. Bioinformatics 26: 976–8.

34. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, et al. (2005) The panther database of protein families, subfamilies, functions and pathways. Nucleic Acids Res 33: D284–8.

35. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, et al. (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 39: D691–7.

36. NCBI Resource Coordinators (2013) Database resources of the national center for biotechnology information. Nucleic Acids Res 41: D8–D20.

37. Desvergne B, Michalik L, Wahli W (2006) Transcriptional regulation of metabolism. Physiol Rev 86: 465–514.

38. Michalik L, Desvergne B, Dreyer C, Gavillet M, Laurini RN, et al. (2002) Ppar expression and function during vertebrate development. Int J Dev Biol 46: 105–14.

39. Thomas PD, Wood V, Mungall CJ, Lewis SE, Blake JA, et al. (2012) On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: A short report. PLoS Comput Biol 8: e1002386.

Figures

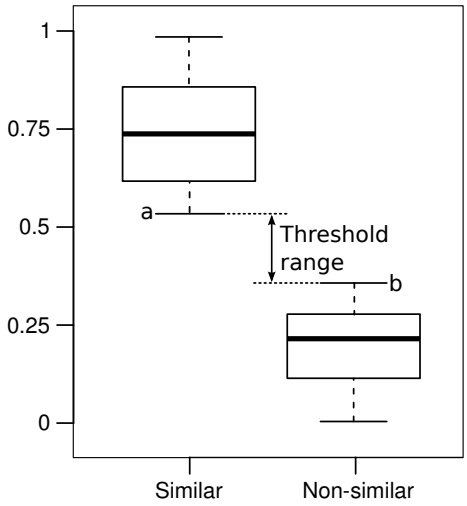


Figure 1. Ideal case of threshold determination. The threshold should be located between the similar and the non-similar boxes.

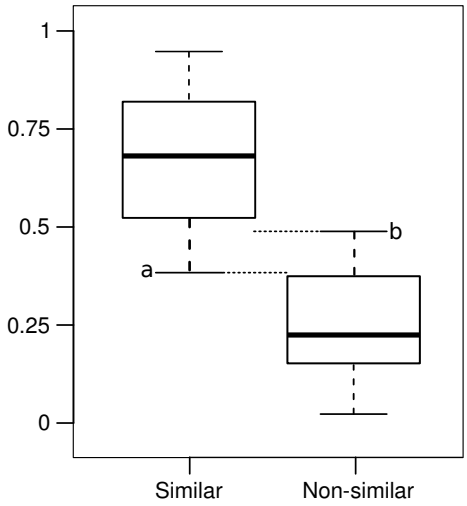


Figure 2. Overlap case in threshold determination. The similar and non-similar boxes overlap.

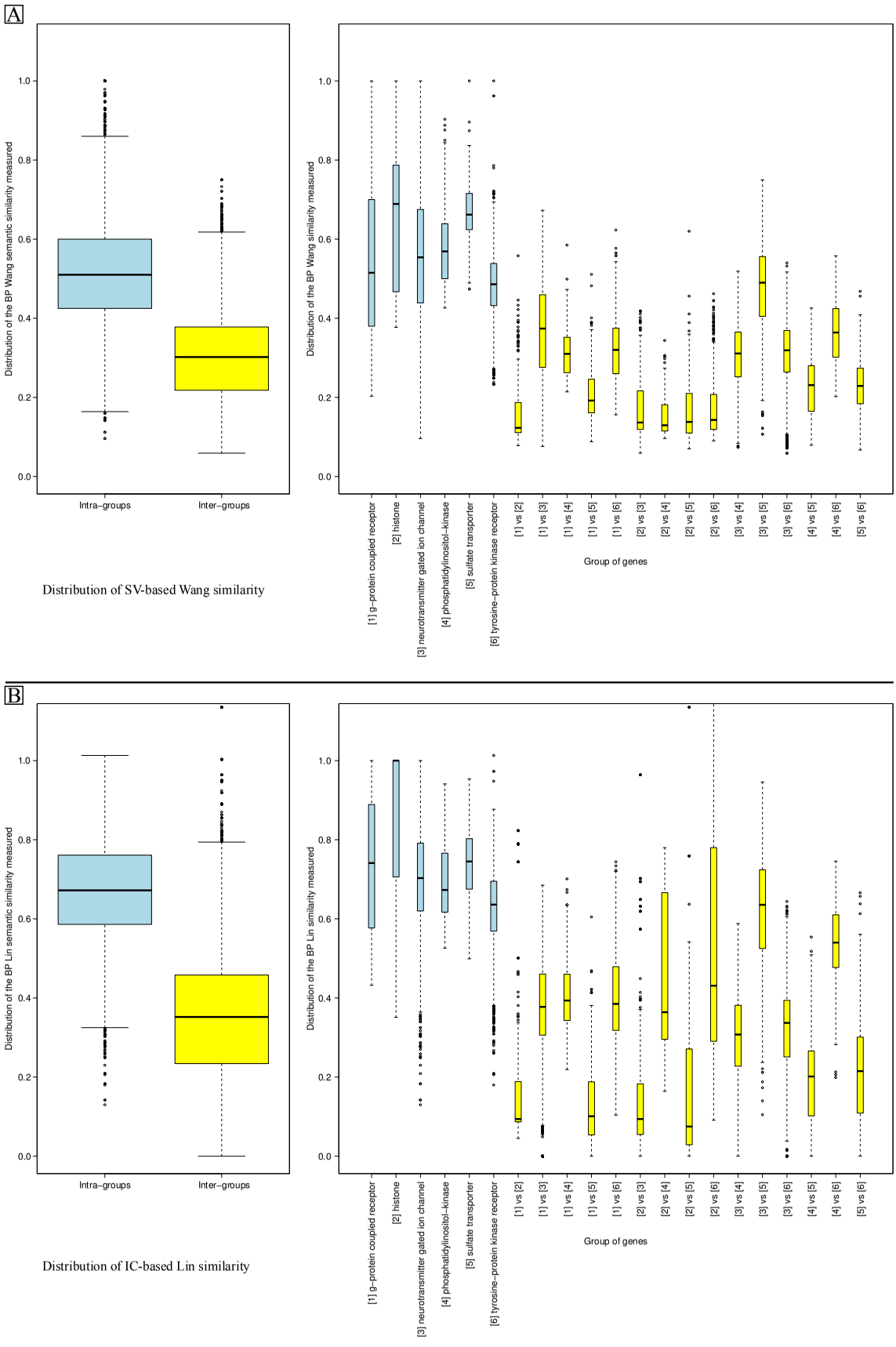


Figure 3. BP distribution of similarity values comparing similar and non-similar genes. Part A concerns the results of the Wang's similarity measure. Part B concerns the results of the Lin's similarity measure.

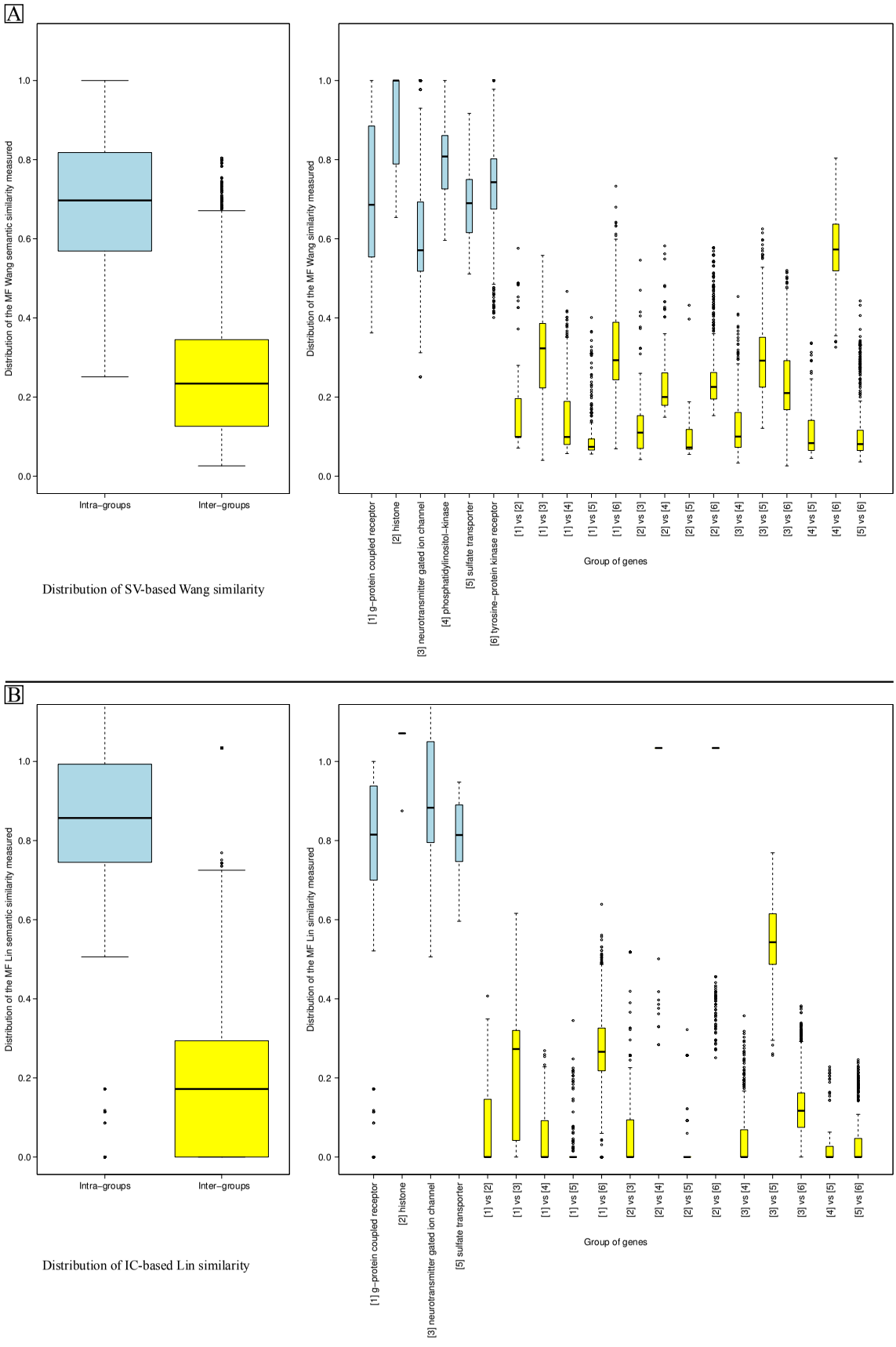


Figure 4. MF distribution of similarity values comparing similar and non-similar genes. Part A concerns the results of the Wang's similarity measure. Part B concerns the results of the Lin's similarity measure.

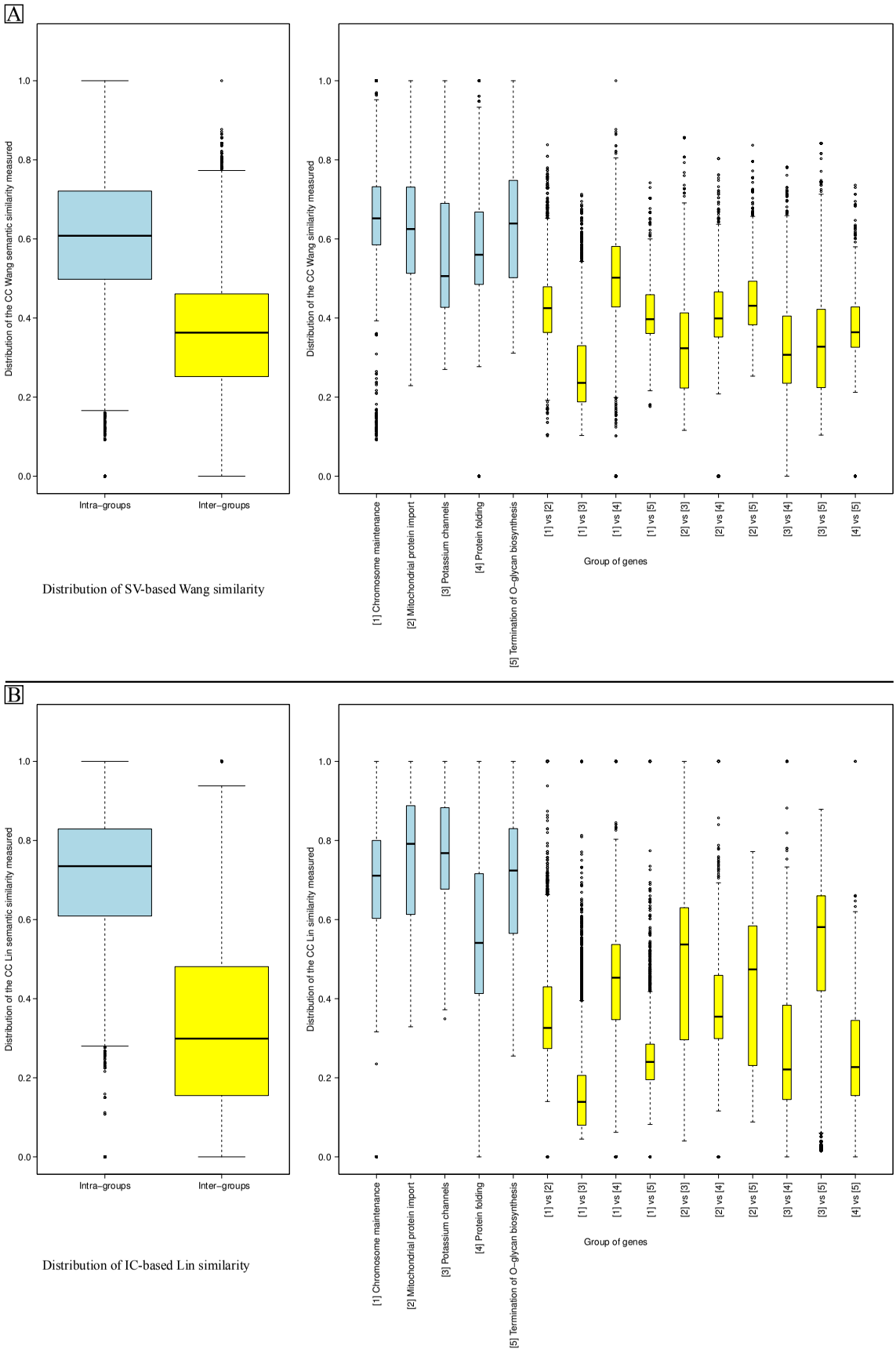


Figure 5. CC distribution of similarity values comparing similar and non-similar genes. Part A concerns the results of the Wang's similarity measure. Part B concerns the results of the Lin's similarity measure.

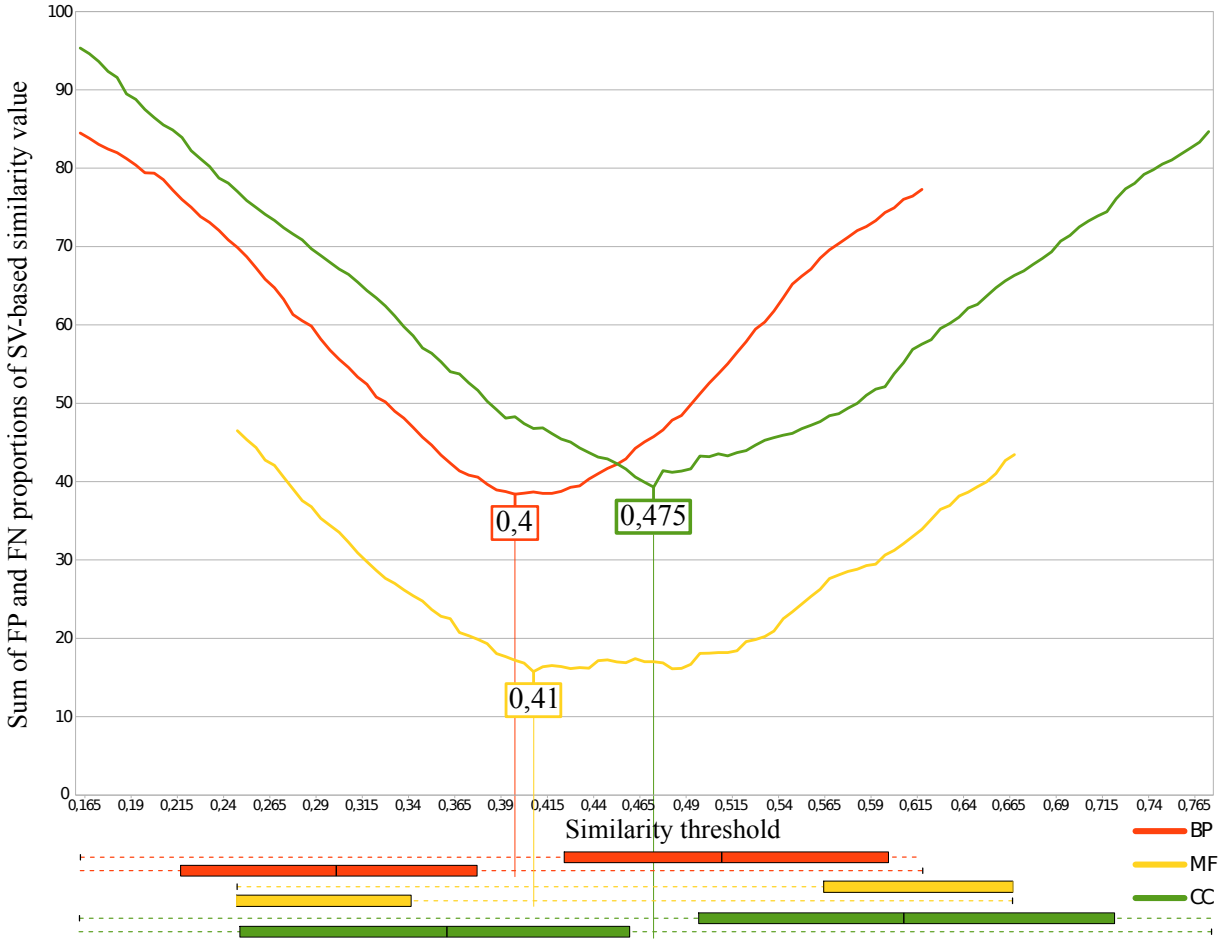


Figure 6. Determination of the Wang's similarity threshold. The minimum of false positive and false negative proportions gives the similarity threshold (τ_{sim}). The overlaping parts of the boxplots (between τ_N and τ_S) from the part A of Figures 3, 4 and 5 are shown in the lower part of the figure. The thresholds are located between the similar and non-similar boxes.

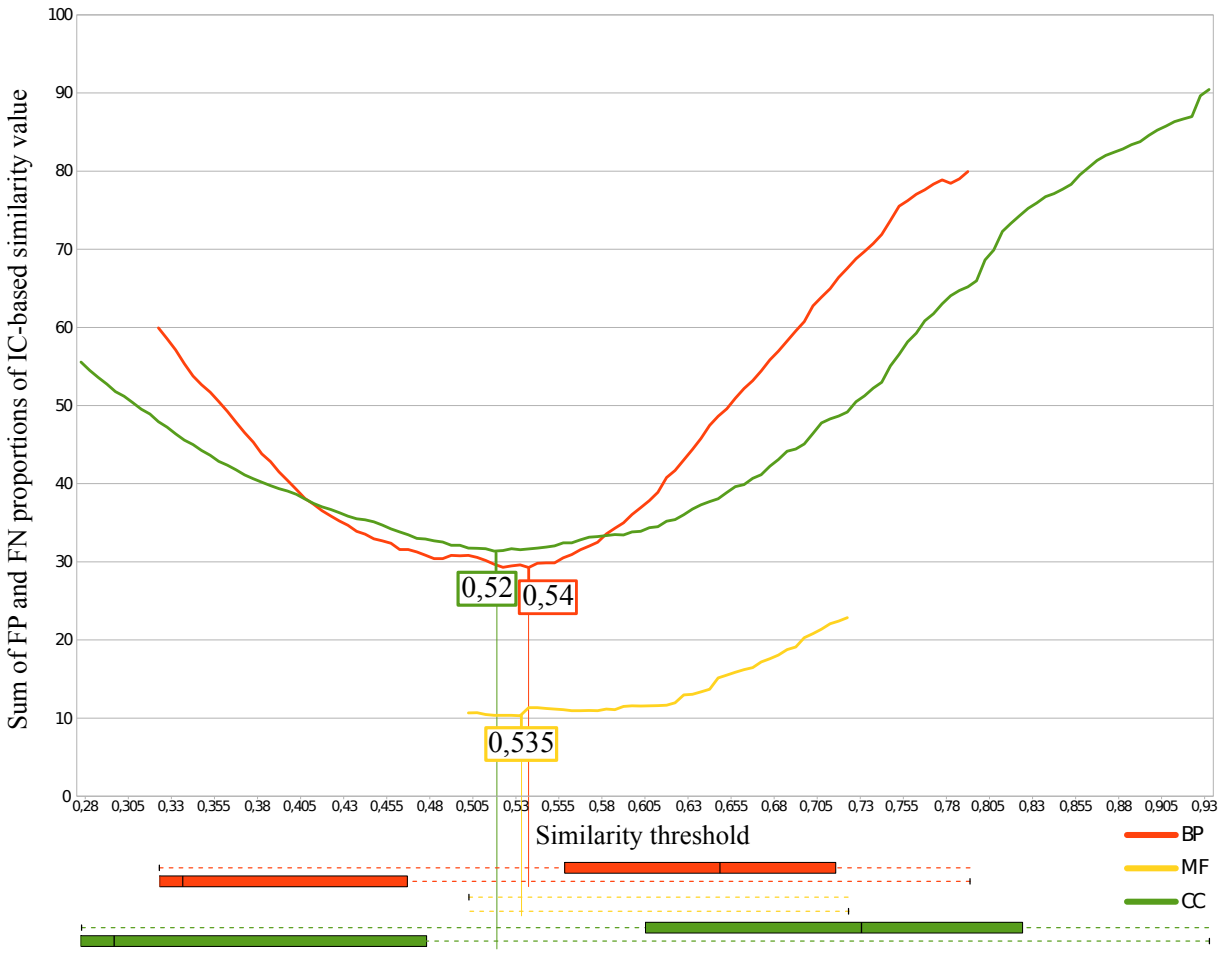


Figure 7. Determination of the Lin's similarity threshold. The minimum of false positive and false negative proportions gives the similarity threshold (τ_{sim}). The overlaping parts of the boxplots (between τ_N and τ_S) from the part B of Figures 3, 4 and 5 are shown in the lower part of the figure. The thresholds are located between the similar and non-similar boxes.

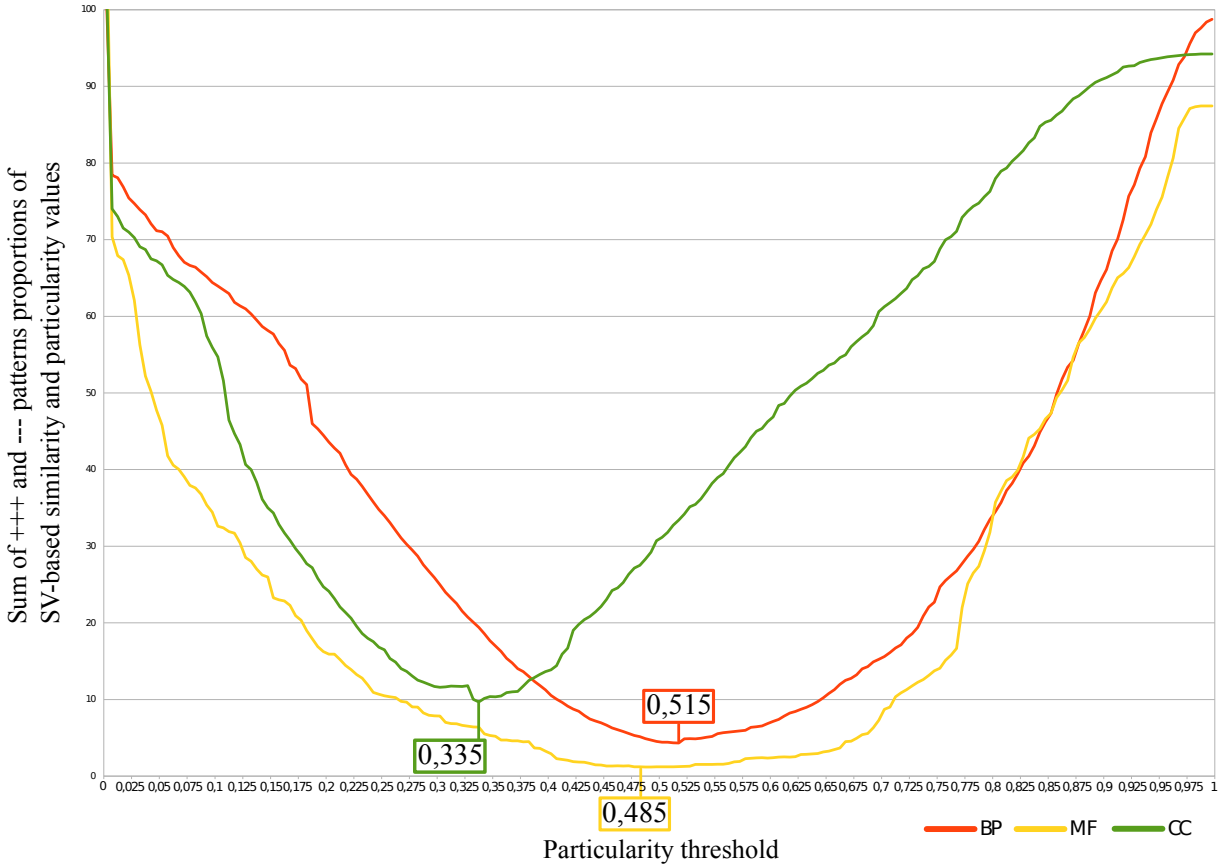


Figure 8. Determination of the SV-based particularity threshold. The minimum of + + + and - - - patterns proportions gives the particularity threshold.

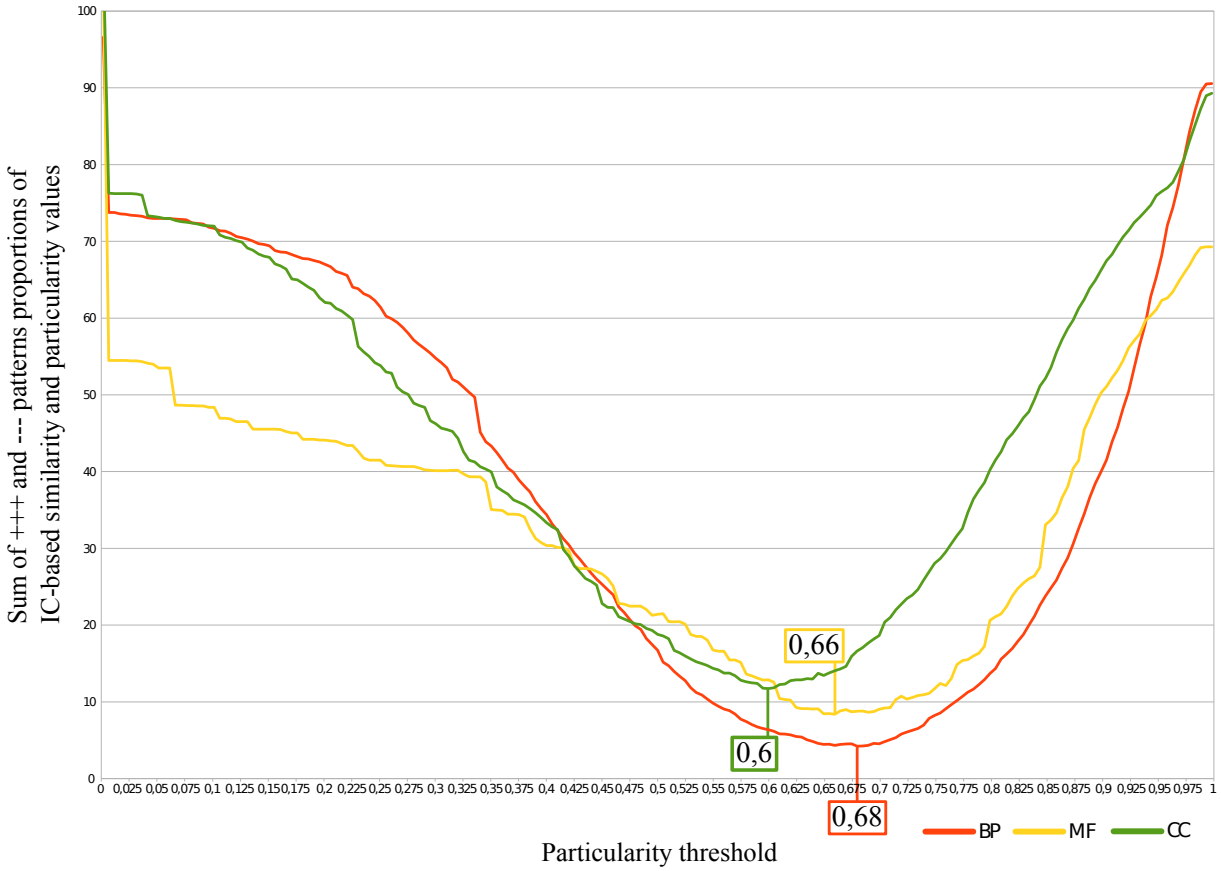


Figure 9. Determination of the IC-based particularity threshold. The minimum of + + + and - - - patterns proportions gives the particularity threshold.

Tables

Table 1. Patterns of similarity and specificity

Notation	sim(A, B)	par(A, B)	par(B, A)
+ + +	$\geqslant \tau_{sim}$	$\geqslant \tau_{par}$	$\geqslant \tau_{par}$
+ + -	$\geqslant \tau_{sim}$	$\geqslant \tau_{par}$	$< \tau_{par}$
+ - +	$\geqslant \tau_{sim}$	$< \tau_{par}$	$\geqslant \tau_{par}$
+ - -	$\geqslant \tau_{sim}$	$< \tau_{par}$	$< \tau_{par}$
- + +	$< \tau_{sim}$	$\geqslant \tau_{par}$	$\geqslant \tau_{par}$
- + -	$< \tau_{sim}$	$\geqslant \tau_{par}$	$< \tau_{par}$
- - +	$< \tau_{sim}$	$< \tau_{par}$	$\geqslant \tau_{par}$
- - -	$< \tau_{sim}$	$< \tau_{par}$	$< \tau_{par}$

The results of a semantic comparison of gene annotations can be summarized in height patterns according to the similarity and particularities values. The first sign is a "+" if the similarity is greater than or equal to the similarity threshold τ_s, a "-" otherwise. The two other signs depends on the two particularity values, a "+" for a particularity greater than the particularity threshold τ_p and a "-" otherwise.

Table 2. Semantic similarity thresholds for Wang and Lin methods

	Wang			Lin		
	τ_N Genes are not similar under	τ_S Genes are similar above	τ_{sim} Threshold minimizing FP and FN	τ_N Genes are not similar under	τ_S Genes are similar above	τ_{sim} Threshold minimizing FP and FN
BP	0.164	0.618	0.4	0.325	0.794	0.54
MF	0.251	0.671	0.41	0.506	0.725	0.535
CC	0.166	0.773	0.475	0.28	0.938	0.52

For each method, τ_N and τ_S respectively give the value of the lowest whisker of the blue box and the value of the upmost whisker of the yellow box of the boxplots in the Figures 3, 4 and 5. For each method, τ_{sim} is the value of threshold that minimizes the proportions of false positive and false negative results, corresponding to the minimum ordinate of the curves in the Figures 6 and 7.

Table 3. Semantic SV-based and IC-based particularity thresholds

	SV-based particularity threshold	IC-based particularity threshold
BP	0.515	0.68
MF	0.485	0.66
CC	0.335	0.6

These thresholds minimize the proportions of non-informative "+ + +" or "- - -" patterns according to Table 1.

Table 4. Thresholds variations considering full and partial datasets (SV-based methods)

Set	Sim. threshold	FN(%)	FP(%)	Par. threshold	+ + + (%)	- - - (%)
BP set	0.4	18.688	19.7	0.515	2.014	2.303
BP set w/o histone	0.42	23.429	16.372	0.495	1.974	2.506
BP set w/o g-protein... receptor	0.405	16.103	17.626	0.515	2.129	2.29
BP set w/o neurotr... channel	0.4	19.276	17.03	0.515	1.715	2.764
BP set w/o tyrosine-... receptor	0.435	27.708	14.451	0.49	1.515	1.648
BP set w/o phosphat...-kinase	0.4	18.954	19.908	0.51	2.154	2.1
BP set w/o sulfate transporter	0.42	23.642	14.784	0.495	1.632	2.624
MF set	0.41	1.602	14.15	0.485	0.894	0.301
MF set w/o histone	0.41	1.625	14.763	0.465	1.121	0.232
MF set w/o g-protein... receptor	0.41	1.831	13.842	0.45	1.795	0.221
MF set w/o neurotr... channel	0.49	4.599	8.668	0.35	0.857	0.599
MF set w/o tyrosine-... receptor	0.41	2.666	12.419	0.485	0.301	0.399
MF set w/o phosphat...-kinase	0.41	1.625	12.666	0.485	1.013	0.253
MF set w/o sulfate transporter	0.41	1.63	14.993	0.485	0.946	0.245
CC set	0.475	17.864	21.443	0.335	5.013	4.677
CC set w/o Chr... maintenance	0.475	27.342	20.251	0.335	5.04	4.583
CC set w/o Mitoch... import	0.475	18.041	21.114	0.335	5.427	3.921
CC set w/o Potassium channels	0.515	15.987	17.133	0.28	7.15	4.902
CC set w/o Protein folding	0.475	17.417	19.082	0.335	4.218	4.97
CC set w/o Term... biosynthesis	0.475	17.867	21.717	0.355	3.873	4.086

This table summarizes the thresholds obtained considering each complete data set or all the groups of a data set except one using Wang similarity measure and SV-based particularity measure. The numbers given for FP, FN, "+ + +" and "- - -" are percentages of the comparison results.

Table 5. Thresholds variations considering full and partial datasets (IC-based methods)

Set	Sim. threshold	FN(%)	FP(%)	Par. threshold	+ + + (%)	- - - (%)
BP set	0.54	16.401	12.88	0.68	1.598	2.665
BP set w/o histone	0.54	16.465	12.326	0.68	1.671	2.896
BP set w/o g-protein... receptor	0.525	14.101	16.081	0.685	1.279	1.71
BP set w/o neurotr... channel	0.525	15.556	15.887	0.69	1.619	1.823
BP set w/o tyrosine-... receptor	0.54	14.403	12.969	0.68	2.168	1.98
BP set w/o phosphat...-kinase	0.525	14.687	14.071	0.685	1.77	2.279
BP set w/o sulfate transporter	0.54	16.633	12.144	0.68	1.484	3.048
MF set	0.535	2.514	7.799	0.66	2.935	5.506
MF set w/o histone	0.535	2.584	5.756	0.66	3.117	5.957
MF set w/o g-protein... receptor	0.565	0.9	9.016	0.66	0.569	1.275
MF set w/o neurotr... channel	0.535	4.258	8.661	0.69	2.98	10.424
MF set w/o tyrosine-... receptor	0.535	2.514	7.849	0.74	1.463	5.248
MF set w/o phosphat...-kinase	0.535	2.514	7.817	0.66	3.015	5.723
MF set w/o sulfate transporter	0.52	2.431	7.265	0.65	2.478	5.956
CC set	0.52	11.838	19.538	0.6	7.155	4.622
CC set w/o Chr... maintenance	0.545	15.222	19.971	0.6	8.717	4.87
CC set w/o Mitoch... import	0.52	12.266	17.596	0.605	7.168	3.983
CC set w/o Potassium channels	0.52	16.347	18.905	0.56	8.427	4.338
CC set w/o Protein folding	0.52	8.072	20.313	0.595	8.026	3.039
CC set w/o Term... biosynthesis	0.52	11.641	18.463	0.6	5.96	4.573

This table summarizes the thresholds obtained considering each complete data set or all the groups of a data set except one using Lin similarity measure and IC-based particularity measure. The numbers given for FP, FN, "+ + +" and "- - -" are percentages of the comparison results.

Table 6. Evolution of patterns in Homologene intra-groups BP comparisons results

BP ↱	+ - -	+ + - or + - +	+ + +	- + +	- + - or - - +	- - -	Total using 0.5 thresholds
+ - -	268,471	0	0	0	0	0	268,471
+ + - or + - +	1,780	54,168	0	0	0	0	55,948
+ + +	7	270	2,623	0	0	0	2,900
- + +	2	154	2,254	10,374	304	1	13,089
- + - or - - +	177	16,027	0	0	32,578	102	48,884
- - -	2,883	0	0	0	0	1,401	4,284
Total using new thresholds	273,320	70,619	4,877	10,374	32,882	1,504	T= 393,576

Numbers of pairs of genes changing from one pattern to another when considering our optimal similarity and particularity thresholds instead of the default values of 0.5. The most important transition consists in 16,027 results moving from the "- + - or - - +" category (size decreased by 32.7%) to the "+ + - or + - +" category (size increased by 26.2%). The number of "+ + +" results is greater with the new thresholds but the number of "- - -" is lower. Globally, the sum of the numbers of the + + + and - - - patterns has decreased (-11.2%).

Table 7. Evolution of patterns in Homologene intra-groups MF comparisons results

MF ↱	+ - -	+ + - or + - +	+ + +	- + +	- + - or - - +	- - -	Total using 0.5 thresholds
+ - -	377,017	2,197	14	0	0	0	379,228
+ + - or + - +	0	37,680	56	0	0	0	37,736
+ + +	0	0	666	0	0	0	666
- + +	0	0	297	8,507	0	0	8,804
- + - or - - +	0	4,738	15	34	12,953	0	17,740
- - -	1,189	87	0	0	25	672	1,973
Total using new thresholds	378,206	44,702	1,048	8,541	12,978	672	T= 446,147

Numbers of pairs of genes changing from one pattern to another when considering our optimal similarity and particularity thresholds instead of the default values of 0.5. After the change of threshold, the most important transition consists in 4,738 results moving from the "- + - or - - +" category (size decreased by 26.8%) to the "+ + - or + - +" category (size increased by 18.5%). The number of "+ + +" results is greater with the new thresholds but the number of "- - -" is lower. Globally, the sum of the numbers of the + + + and - - - patterns has decreased (-34.8%).

Table 8. Evolution of patterns in Homologene intra-groups CC comparisons results

CC ↱	+ - -	+ + - or + - +	+ + +	- + +	- + - or - - +	- - -	Total using 0.5 thresholds
+ - -	250,826	25,089	948	0	0	0	276,863
+ + - or + - +	0	67,349	2,103	0	0	0	69,452
+ + +	0	0	1,237	0	0	0	1,237
- + +	0	0	104	2,746	0	0	2,850
- + - or - - +	0	2,292	90	1,191	19,956	0	23,529
- - -	118	196	34	69	470	369	1,256
Total using new thresholds	250,944	94,926	4,516	4,006	20,426	369	T= 375,187

Numbers of pairs of genes changing from one pattern to another when considering our optimal similarity and particularity thresholds instead of the default values of 0.5. After the change of threshold, the most important transition consists in 25,089 results moving from the "+ - -" category (size decreased by 9.4%) to the "+ + - or + - +" category (size increased by 36.7%). The number of "+ + +" results is greater with the new thresholds but the number of "- - -" is lower. Globally, the sum of the numbers of the + + + and - - - patterns has increased (+49%).

Table 9. SV-based BP similarity and particularity measured between orthologs and paralogs of the PPAR family

BP			α mmu	α rno	α mamu	α hsa	α cca	α bta	β mmu	β rno	β mamu	β hsa	β cca	β bta	γ mmu	γ rno	γ mamu	γ hsa	γ cca	γ bta
		Annot	32	30	20	39	21	21	30	49	25	37	25	25	61	56	38	65	38	38
SIM	α mmu	32	1	0.983	0.847	0.946	0.859	0.859	0.608	0.601	0.611	0.628	0.611	0.611	0.642	0.654	0.616	0.65	0.616	0.616
	α rno	30		1	0.869	0.945	0.877	0.877	0.598	0.592	0.6	0.598	0.6	0.6	0.645	0.657	0.628	0.644	0.628	0.628
	α mamu	20			1	0.814	0.993	0.993	0.625	0.563	0.634	0.612	0.634	0.634	0.606	0.614	0.627	0.601	0.627	0.627
	α hsa	39				1	0.831	0.831	0.634	0.623	0.637	0.655	0.637	0.637	0.663	0.677	0.634	0.694	0.634	0.634
	α cca	21					1	1	0.63	0.578	0.64	0.626	0.64	0.64	0.612	0.625	0.621	0.608	0.621	0.621
	α bta	21						1	0.63	0.578	0.64	0.626	0.64	0.64	0.612	0.625	0.621	0.608	0.621	0.621
	β mmu	30							1	0.83	0.948	0.917	0.948	0.948	0.647	0.643	0.663	0.638	0.663	0.663
	β rno	49								1	0.823	0.822	0.823	0.823	0.715	0.714	0.644	0.711	0.644	0.644
	β mamu	25									1	0.929	1	1	0.642	0.65	0.68	0.644	0.68	0.68
	β hsa	37										1	0.929	0.929	0.642	0.652	0.662	0.661	0.662	0.662
	β cca	25											1	1	0.642	0.65	0.68	0.644	0.68	0.68
	β bta	25												1	0.642	0.65	0.68	0.644	0.68	0.68
	γ mmu	61													1	0.978	0.882	0.959	0.882	0.882
	γ rno	56														1	0.896	0.97	0.896	0.896
	γ mamu	38															1	0.868	1	1
	γ hsa	65																1	0.868	0.868
	γ cca	38																	1	1
	γ bta	38																		1
PAR	α mmu	32	0	0	0.112	0	0.112	0.112	0.481	0.442	0.487	0.474	0.487	0.487	0.312	0.315	0.437	0.308	0.437	0.437
	α rno	30	0	0	0.112	0	0.112	0.112	0.481	0.442	0.487	0.474	0.487	0.487	0.312	0.315	0.437	0.308	0.437	0.437
	α mamu	20	0	0	0	0	0	0	0.424	0.408	0.429	0.414	0.429	0.429	0.282	0.285	0.405	0.277	0.405	0.405
	α hsa	39	0.167	0.167	0.259	0	0.259	0.259	0.502	0.469	0.507	0.483	0.507	0.507	0.398	0.4	0.502	0.362	0.502	0.502
	α cca	21	0	0	0	0	0	0	0.424	0.408	0.429	0.414	0.429	0.429	0.282	0.285	0.405	0.277	0.405	0.405
	α bta	21	0	0	0	0	0	0	0.424	0.408	0.429	0.414	0.429	0.429	0.282	0.285	0.405	0.277	0.405	0.405
	β mmu	30	0.441	0.441	0.449	0.357	0.449	0.449	0	0.009	0.144	0.127	0.144	0.144	0.265	0.364	0.412	0.361	0.412	0.412
	β rno	49	0.603	0.603	0.626	0.548	0.626	0.626	0.346	0	0.435	0.405	0.435	0.435	0.424	0.491	0.578	0.489	0.578	0.578
	β mamu	25	0.355	0.355	0.362	0.256	0.362	0.362	0	0	0	0	0	0	0.27	0.27	0.327	0.27	0.327	0.327
	β hsa	37	0.417	0.417	0.423	0.313	0.423	0.423	0.101	0.073	0.119	0	0.119	0.119	0.341	0.341	0.391	0.325	0.391	0.391
	β cca	25	0.355	0.355	0.362	0.256	0.362	0.362	0	0	0	0	0	0	0.27	0.27	0.327	0.27	0.327	0.327
	β bta	25	0.355	0.355	0.362	0.256	0.362	0.362	0	0	0	0	0	0	0.27	0.27	0.327	0.27	0.327	0.327
	γ mmu	61	0.548	0.548	0.581	0.526	0.581	0.581	0.551	0.467	0.619	0.61	0.619	0.619	0	0.104	0.32	0.102	0.32	0.32
	γ rno	56	0.498	0.498	0.534	0.473	0.534	0.534	0.567	0.475	0.575	0.565	0.575	0.575	0	0	0.241	0	0.241	0.241
	γ mamu	38	0.456	0.456	0.489	0.422	0.489	0.489	0.473	0.427	0.483	0.47	0.483	0.483	0	0	0	0	0	0
	γ hsa	65	0.52	0.52	0.554	0.469	0.554	0.554	0.588	0.501	0.597	0.577	0.597	0.597	0.051	0.053	0.282	0	0.282	0.282
	γ cca	38	0.456	0.456	0.489	0.422	0.489	0.489	0.473	0.427	0.483	0.47	0.483	0.483	0	0	0	0	0	0
	γ bta	38	0.456	0.456	0.489	0.422	0.489	0.489	0.473	0.427	0.483	0.47	0.483	0.483	0	0	0	0	0	0

Green cells contain similarity values greater than τ_S, red cells contain similarity values lower than τ_S, yellow cells contain values greater than τ_{par} and blue cells contain values lower than τ_{par}. All orthologs have a "+ - -" pattern and some paralogs have a "- + -" or a "+ + -" pattern.

Table 10. SV-based MF similarity and particularity measured between orthologs and paralogs of the PPAR family

MF			α mmu	α rno	α mamu	α hsa	α cca	α bta	β mmu	β rno	β mamu	β hsa	β cca	β bta	γ mmu	γ rno	γ mamu	γ hsa	γ cca	γ bta
		Annot	19	18	12	18	12	12	14	13	10	11	10	10	17	16	14	18	14	14
SIM	α mmu	19	1	0.992	0.89	0.992	0.89	0.89	0.8	0.794	0.79	0.782	0.79	0.79	0.824	0.836	0.811	0.82	0.811	0.811
	α rno	18		1	0.906	1	0.906	0.906	0.784	0.805	0.804	0.795	0.804	0.804	0.827	0.841	0.819	0.823	0.819	0.819
	α mamu	12			1	0.906	1	1	0.822	0.851	0.893	0.879	0.893	0.893	0.8	0.815	0.826	0.796	0.826	0.826
	α hsa	18				1	0.906	0.906	0.784	0.805	0.804	0.795	0.804	0.804	0.827	0.841	0.819	0.823	0.819	0.819
	α cca	12					1	1	0.822	0.851	0.893	0.879	0.893	0.893	0.8	0.815	0.826	0.796	0.826	0.826
	α bta	12						1	0.822	0.851	0.893	0.879	0.893	0.893	0.8	0.815	0.826	0.796	0.826	0.826
	β mmu	14							1	0.859	0.893	0.89	0.893	0.893	0.814	0.807	0.809	0.835	0.809	0.809
	β rno	13								1	0.931	0.941	0.931	0.931	0.851	0.831	0.84	0.844	0.84	0.84
	β mamu	10									1	0.992	1	1	0.809	0.8	0.812	0.805	0.812	0.812
	β hsa	11										1	0.992	0.992	0.806	0.784	0.794	0.803	0.794	0.794
	β cca	10											1	1	0.809	0.8	0.812	0.805	0.812	0.812
	β bta	10												1	0.809	0.8	0.812	0.805	0.812	0.812
	γ mmu	17													1	0.978	0.967	0.986	0.967	0.967
	γ rno	16														1	0.991	0.965	0.99	0.991
	γ mamu	14															1	0.954	1	1
	γ hsa	18																1	0.954	0.954
	γ cca	14																	1	1
	γ bta	14																		1
PAR	α mmu	19	0	0.054	0.188	0.054	0.188	0.188	0.315	0.369	0.416	0.416	0.416	0.416	0.344	0.365	0.365	0.344	0.365	0.365
	α rno	18	0	0	0.141	0	0.141	0.141	0.333	0.333	0.382	0.382	0.382	0.382	0.307	0.329	0.329	0.307	0.329	0.329
	α mamu	12	0	0	0	0	0	0	0.224	0.224	0.281	0.281	0.281	0.281	0.193	0.218	0.218	0.193	0.218	0.218
	α hsa	18	0	0	0.141	0	0.141	0.141	0.333	0.333	0.382	0.382	0.382	0.382	0.307	0.329	0.329	0.307	0.329	0.329
	α cca	12	0	0	0	0	0	0	0.224	0.224	0.281	0.281	0.281	0.281	0.193	0.218	0.218	0.193	0.218	0.218
	α bta	12	0	0	0	0	0	0	0.224	0.224	0.281	0.281	0.281	0.281	0.193	0.218	0.218	0.193	0.218	0.218
	β mmu	14	0.492	0.532	0.532	0.532	0.532	0.532	0	0.368	0.402	0.402	0.402	0.402	0.27	0.417	0.417	0.13	0.417	0.417
	β rno	13	0.35	0.35	0.35	0.35	0.35	0.35	0.123	0	0.17	0.17	0.17	0.17	0.076	0.279	0.279	0.076	0.279	0.279
	β mamu	10	0.274	0.274	0.274	0.274	0.274	0.274	0	0	0	0	0	0	0.055	0.3	0.3	0.055	0.3	0.3
	β hsa	11	0.274	0.274	0.274	0.274	0.274	0.274	0	0	0	0	0	0	0.055	0.3	0.3	0.055	0.3	0.3
	β cca	10	0.274	0.274	0.274	0.274	0.274	0.274	0	0	0	0	0	0	0.055	0.3	0.3	0.055	0.3	0.3
	β bta	10	0.274	0.274	0.274	0.274	0.274	0.274	0	0	0	0	0	0	0.055	0.3	0.3	0.055	0.3	0.3
	γ mmu	17	0.568	0.568	0.568	0.568	0.568	0.568	0.352	0.408	0.498	0.498	0.498	0.498	0	0.244	0.244	0	0.244	0.244
	γ rno	16	0.446	0.446	0.446	0.446	0.446	0.446	0.314	0.389	0.508	0.508	0.508	0.508	0	0	0	0	0	0
	γ mamu	14	0.446	0.446	0.446	0.446	0.446	0.446	0.314	0.389	0.508	0.508	0.508	0.508	0	0	0	0	0	0
	γ hsa	18	0.615	0.615	0.615	0.615	0.615	0.615	0.313	0.474	0.554	0.554	0.554	0.554	0.111	0.328	0.328	0	0.328	0.328
	γ cca	14	0.446	0.446	0.446	0.446	0.446	0.446	0.314	0.389	0.508	0.508	0.508	0.508	0	0	0	0	0	0
	γ bta	14	0.446	0.446	0.446	0.446	0.446	0.446	0.314	0.389	0.508	0.508	0.508	0.508	0	0	0	0	0	0

Green cells contain similarity values greater than τ_S, red cells contain similarity values lower than τ_S, yellow cells contain values greater than τ_{par} and blue cells contain values lower than τ_{par}. All orthologs have a "+ - -" pattern and some paralogs have a "+ + -" pattern.

3 Synthèse

La définition de seuils de similarité pour différentes mesures couramment utilisées permet d'identifier les produits de gènes similaires. Le seuil défini pour la méthode de Wang est utile pour identifier des orthologues intervenant dans les voies métaboliques homologues de différentes espèces comme similaires. D'après les résultats obtenus à partir de la base de données HomoloGene, la plupart des orthologues correctement annotés sont similaires. La définition du seuil de particularité permet de savoir si les fonctions spécifiques à un produit de gène lorsqu'on le compare à un produit de gène similaire d'une autre espèce sont anecdotiques ou importantes. Ces seuils nous permettent l'interprétation des résultats obtenus lors de la comparaison inter-espèces systématique de tous les produits de gènes d'une voie métabolique homologue.

CHAPITRE 5

COMPARAISON INTER-ESPÈCES DU MÉTABOLISME DES LIPIDES

Dans ce chapitre, nous appliquons les mesures de similarité et particularité présentées dans les chapitres précédents aux gènes impliqués dans le métabolisme des lipides chez l'Homme, la souris et la poule. Nous utilisons les seuils définis dans le chapitre 4 pour interpréter les résultats de ces comparaisons.

Sommaire

1 Comparaison structurelle d'une voie métabolique entre 2 espèces

La méthodologie décrite dans les deux chapitres précédents permet de comparer des gènes sur la base de leurs annotations fonctionnelles. Les produits de gènes permettent le fonctionnement des voies métaboliques, notamment en catalysant les réactions biochimiques. La comparaison inter-espèces de voies métaboliques doit commencer par l'identification de réactions communes aux espèces que l'on souhaite comparer. Ensuite, il est possible de réaliser la comparaison sémantique des gènes impliqués dans ces réactions.

Parmi les grandes bases de données présentées dans le chapitre 2, Reactome fournit les données concernant le métabolisme d'une vingtaine d'espèces dont *Homo sapiens*, *Mus musculus* et *Gallus gallus*, au format BioPAX v3 (un fichier RDF/OWL par espèce). Ce format est un des standards principaux pour représenter les voies métaboliques. Notre approche repose sur ces données, mais peut être appliquée à d'autres sources que Reactome puisqu'une étude récente a mis en évidence la complémentarité des bases de voies métaboliques [Soh et al., 2010].

La Figure 1 présente la hierarchie des classes servant à décrire les voies métaboliques d'une espèce dans un fichier BioPAX de Reactome. Chaque classe est décrite par des propriétés. Chaque instance de la classe "*Pathway*" a ainsi une propriété "*pathwayComponent*" qui fait la liste de toutes les subdivisions d'une voie métabolique. Ces subdivisions sont à leur tour soit des instances de la classe "*Pathway*", soit des instances de la classe "*BiochemicalReaction*".

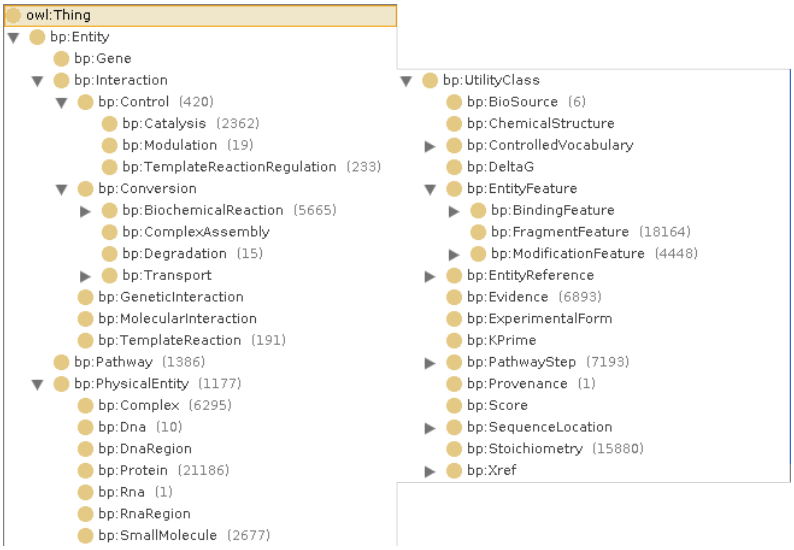


FIGURE 1 – Hiérarchie des classes présentes dans un fichier BioPAX de Reactome. Le nombre d'instances des différentes classes pour *Mus musculus* figure entre parenthèses à côté du nom des classes.

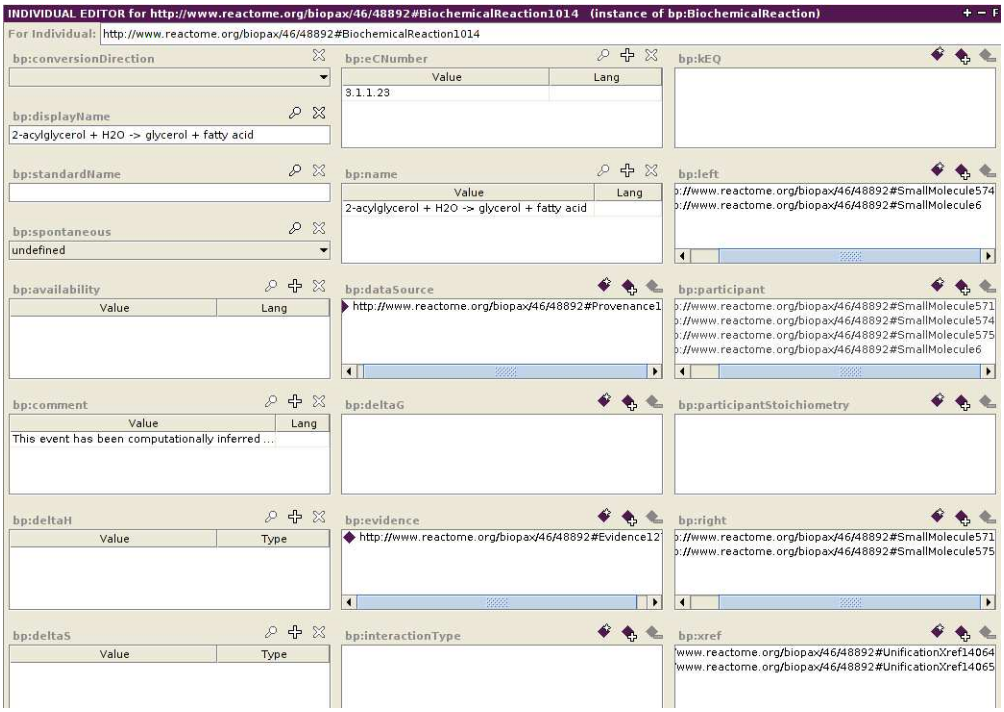


FIGURE 2 – L'instance "*#BiochemicalReaction1014*" de la classe "*BiochemicalReaction*" du fichier concernant le métabolisme de *Mus musculus* issu de Reactome. Toutes les données ont un identifiant au sein du fichier permettant de naviguer entre les instances des différentes classes.

Ainsi, chez *Homo sapiens*, le "*Pathway*" ayant pour nom "*Metabolism of lipids and lipoproteins*" a 12 "*pathwayComponent*" qui sont eux-même des instances de la classe "*Pathway*". Un de ces "*Pathway*"est "*Cholesterol biosynthesis*" dont les 25 "*pathwayComponent*" sont tous des instances de "*BiochemicalReaction*". Cette organisation permet de décrire finement une voie métabolique.

La Figure 2 montre les propriétés de l'instance "*#BiochemicalReaction1014*" de *Mus musculus*. Grâce aux propriétés, on sait que cette instance est nommée d'après la réaction qu'elle décrit. En effet, la propriété "*displayName*" est : « 2-acylglycerol + H2O → glycerol + fatty acid ». On sait que cette réaction est catalysée par une enzyme ayant pour code EC 3.1.1.23 et on connaît les molécules qui participent à la réaction ("*participant*") en étant capable de distinguer les substrats ("*left*") des produits ("*right*"). Seuls les identifiants internes au fichier de Reactome représentent ici les participants, leur description complète étant disponible dans les instances correspondantes des sous-classes de "*PhysicalEntity*" (dans cet exemple, "*SmallMolecule*", mais il peut également s'agir de "*Complex*" et de "*Protein*").

Il est possible de représenter les instances listées dans la propriété "*pathwayComponent*" de l'instance nommée "Metabolism" sous la forme d'un graphe. En procédant à une expansion systématique de cette propriété, on obtient le graphe complet du métabolisme pour une espèce. L'instance "Metabolism" de Reactome rassemble les mécanismes d'anabolisme (synthèse) et de catabolisme (dégradation) décrits dans la Figure 3.

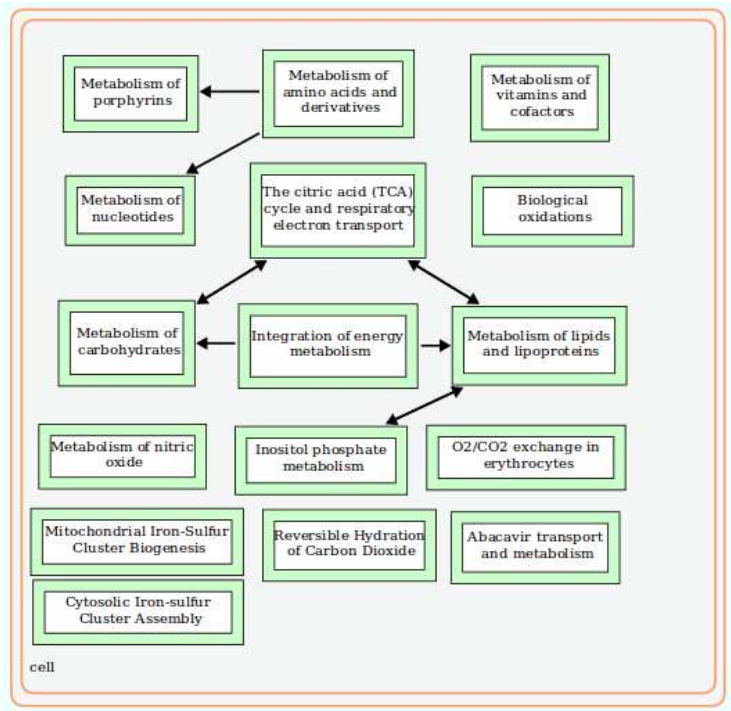


FIGURE 3 – Instances de "*Pathway*" attachées à "Metabolism" dans Reactome.

La comparaison des fichiers de plusieurs espèces permet d'identifier les instances communes et celles qui sont spécifiques d'une espèce. Cela permet la réalisation de graphes inter-espèces, comme celui présenté dans la Figure 4, qui représente la comparaison structurelle des voies métaboliques d'*Homo sapiens* et de *Mus musculus* classées sous l'instance "Metabolism".

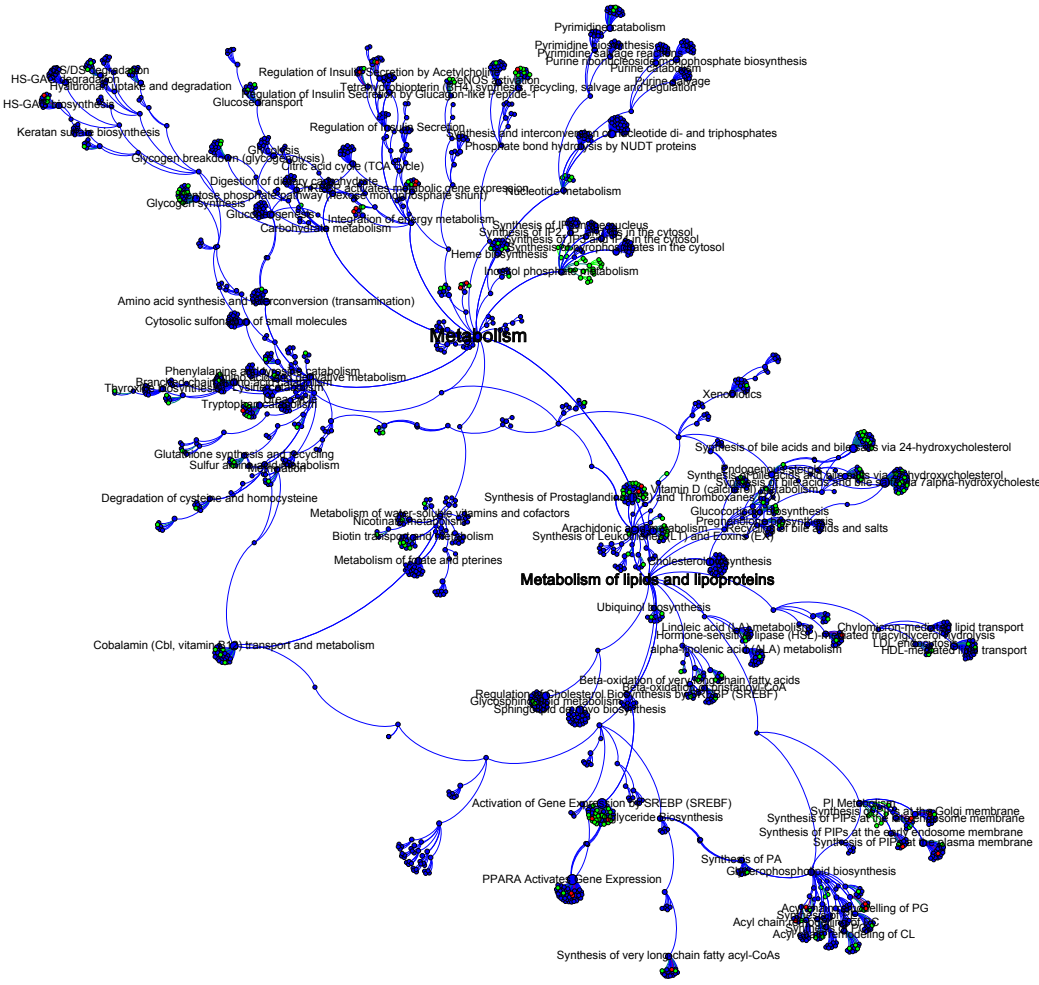


FIGURE 4 – Graphe complet de la comparaison structurelle du métabolisme de l'Homme et de la souris. Les étapes communes sont en bleu, les étapes spécifiques à l'Homme sont en vert et celles spécifiques à la souris en rouge.

Dans ce graphe, les nœuds centraux appartiennent à la classe "*Pathway*" et les feuilles à la classe "*BiochemicalReaction*". On peut voir que sur la totalité du métabolisme, toutes les instances des classes "*Pathway*" sont communes à l'Homme et à la souris. En effet, tous les nœuds sont bleus et seules une minorité de feuilles sont vertes (indiquant une réaction décrite seulement chez l'Homme) et encore moins sont rouges (indiquant une réaction décrite seulement chez la souris). Il n'est pas suprenant d'observer une structure si proche entre le métabolisme de deux mammifères. Cependant il faut garder à l'esprit que Reactome rassemble avant tout la connaissance sur les voies métaboliques humaines,

les données disponibles pour les autres espèces étant produites par une inférence dont le résultat est vérifié manuellement. Comme les données d'annotations présentées et utilisées dans les chapitres précédents, les données qui décrivent les voies métaboliques reflètent la connaissance disponible. Le fait qu'une réaction apparaisse comme spécifique à une espèce ne veut pas nécessairement dire que celle-ci n'existe pas chez l'autre, mais qu'on n'a pas connaissance de son existence.

Les Figures 5 et 6 détaillent les points communs et différences dans la structure du métabolisme des lipides respectivement chez *Homo sapiens* et *Mus musculus* et chez *Homo sapiens* et *Gallus gallus*. Les nœuds présents chez les deux espèces comparées sont en bleu, ceux spécifiques à l'Homme en vert et ceux spécifique à la souris ou à la poule, en rouge.

Dans les deux comparaisons, toutes les instances de "*Pathway*" sont communes aux deux espèces comparées. Les différences que l'on observe sont au niveau des feuilles (réactions). L'Homme a plus de réactions spécifiques que la souris et la poule.

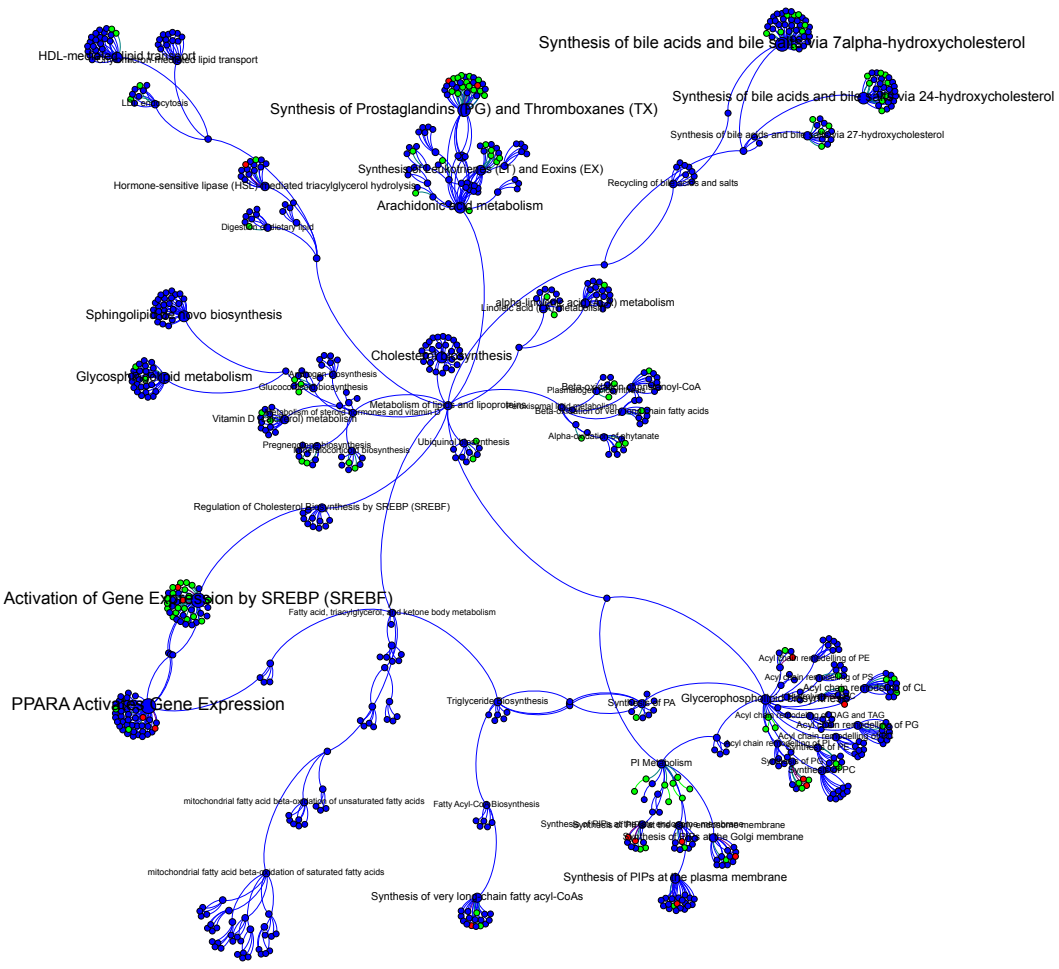


FIGURE 5 – Graphe de la comparaison structurelle du métabolisme des lipides de l'Homme et de la souris. Les étapes communes sont en bleu, les étapes spécifiques à l'Homme sont en vert et celles spécifiques à la souris en rouge.

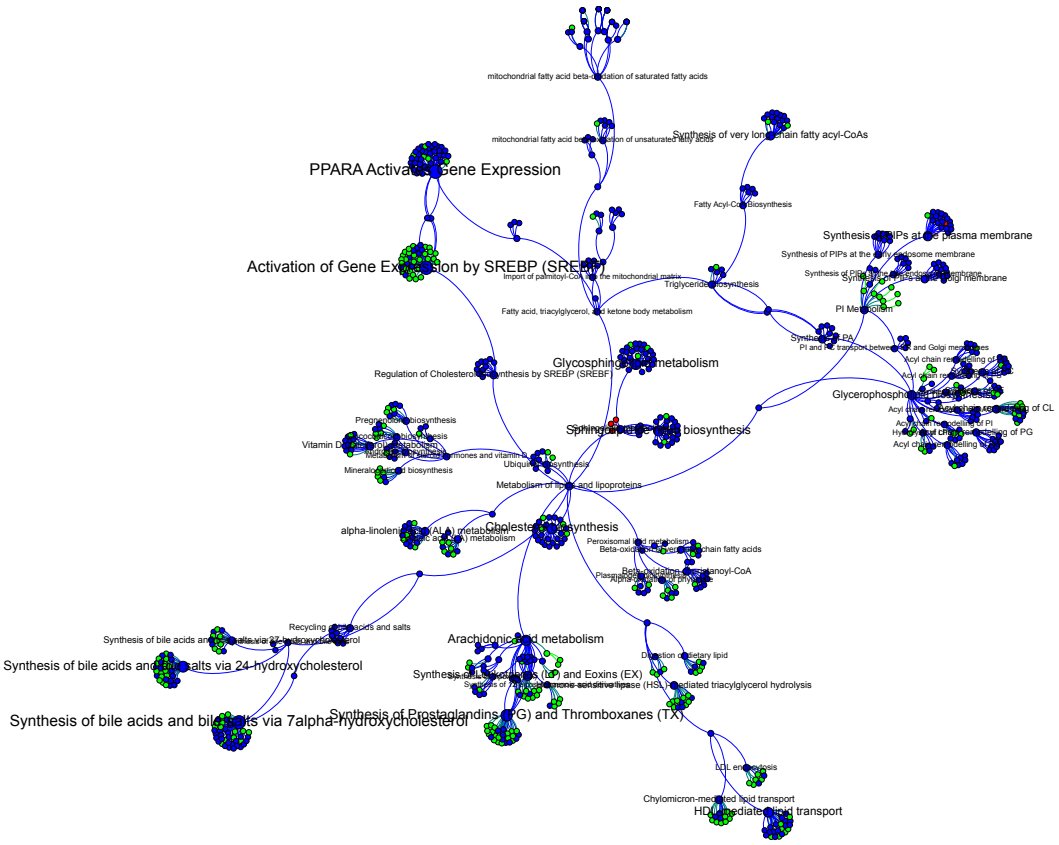


FIGURE 6 – Graphe de la comparaison structurelle du métabolisme des lipides de l'Homme et de la poule. Les étapes communes sont en bleu, les étapes spécifiques à l'Homme sont en vert et celles spécifiques à la poule en rouge.

En effet, lorsque l'on compare l'Homme et la souris, on trouve 650 réactions communes, 119 réactions spécifiques à l'Homme et 17 réactions spécifiques à la souris. Et lorsque l'on compare l'Homme et la poule, on trouve 556 réactions communes, 233 réactions spécifiques à l'Homme et 4 réactions spécifiques à la poule.

Si l'on regarde en détail les 4 réactions marquées comme spécifiques de la poule, on remarque qu'il s'agit vraisemblablement d'erreurs. En effet, dans la Figure 7, on peut voir que deux de ces réactions existent aussi au niveau de feuilles marquées communes via l'utilisation d'un synonyme : la « 3-dehydrosphinganine » est aussi appelée « 3-ketosphinganine ». La troisième réaction marquée comme spécifique à la poule sur cette figure diffère d'une réaction marquée comme commune par le sens de la réaction. Le signe "$\rightarrow$" est utilisé pour la feuille rouge alors que "$\rightleftarrows$" est présent au niveau de la feuille bleue. Ces trois feuilles marquées comme spécifiques de la poule sont directement attachées au terme général "*Sphingolipid metabolism*" alors que les feuilles marquées comme communes correspondantes dépendent d'un niveau plus précis : "*Sphingolipid de novo biosynthesis*". Enfin, la Figure 8 montre que le nom de la dernière réaction marquée comme spécifique de la poule est une version plus générale du nom d'une réaction marquée comme spécifique de l'Homme. Il devrait donc ici aussi s'agir d'une réaction commune.

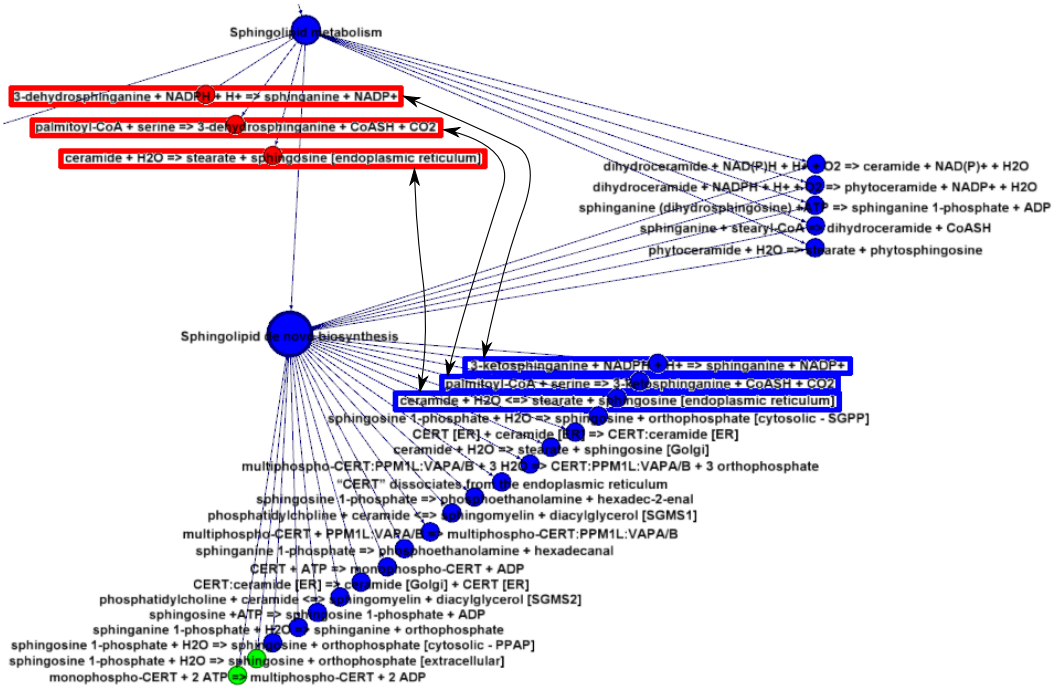


FIGURE 7 – Réactions marquées à tort comme spécifiques de la poule (1). Les feuilles rouges sont spécifiques de la poule, mais les réactions qu'elles représentent existent également au niveau de feuilles bleues (communes).

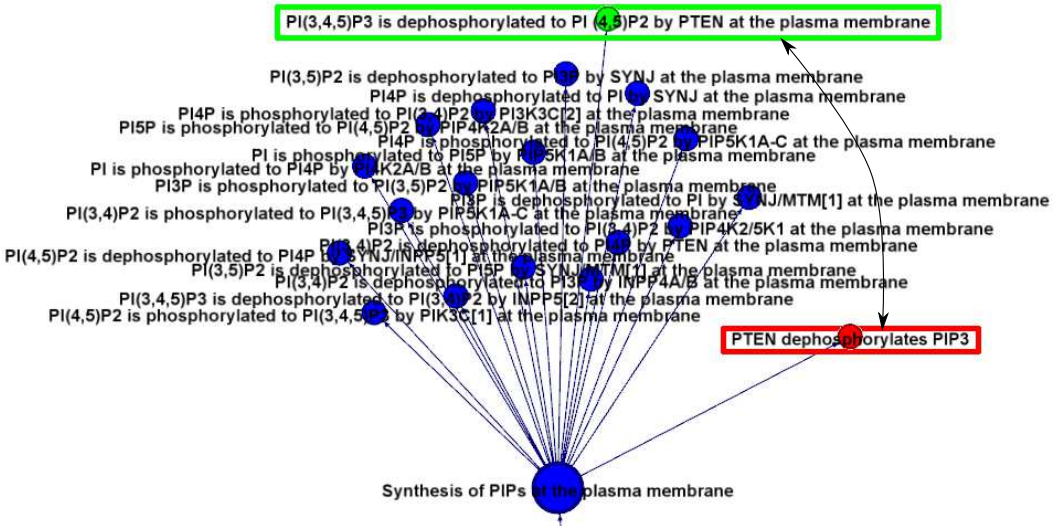


FIGURE 8 – Réaction marquée à tort comme spécifique de la poule (2). La feuille rouge est spécifique de la poule, mais la réaction qu'elle représente existe également au niveau d'une feuille verte (spécifique de l'Homme). Par conséquent, ces deux feuilles devraient être remplacées par une seule feuille bleue.

2 Comparaison fonctionnelle du métabolisme des lipides chez l'humain, la souris et la poule

Parmi les propriétés des instances de la classe "*BiochemicalReaction*", Reactome mentionne le code EC de l'enzyme qui catalyse la réaction. Il n'y a généralement qu'un code EC associé à une réaction, mais il est possible, bien que rare, d'avoir plusieurs enzymes impliqués dans une même réaction. Une même enzyme peut intervenir dans plusieurs réactions. Il est également possible qu'une réaction soit spontanée et ne nécessite pas d'enzyme. Plusieurs produits de gènes sont souvent associés à un code EC. Il s'agit dans ce cas d'isozymes, c'est-à-dire de produits de gènes différents catalysant une même réaction. Ce système permet une meilleure plasticité du métabolisme, certaines isozymes étant plus efficaces dans une condition donnée ou dans un tissu particulier. Pour chaque réaction classée comme commune à l'étape de comparaison structurelle décrite ci-dessus, on peut donc établir une liste de produit de gènes associés pour les 2 espèces à comparer.

L'utilisation d'une mesure de similarité sémantique et de notre mesure de particularité sémantique permet de comparer les gènes de deux espèces associés à une même réaction biochimique. Puisqu'il s'agit d'une comparaison inter-espèces, nous avons choisi d'appliquer la mesure de similarité de Wang et de baser notre mesure de particularité sur la valeur sémantique des termes GO qui annotent les gènes à comparer.

Notre approche procède en deux temps. Premièrement, la combinaison des mesures de similarité et de particularité dont l'intérêt a été présenté dans le chapitre 3. Deuxièmement, l'utilisation des seuils de similarité et de particularité définis au chapitre 4 pour faciliter l'interprétation des résultats.

Nous présentons ici les résultats obtenus par cette approche appliquée aux produits de gènes qui interviennent dans les réactions du métabolisme des lipides communes à l'Homme et la souris, ainsi qu'à l'Homme et la poule. La représentation des résultats de la comparaison fonctionnelle sur les graphes de comparaison du métabolisme permet de rassembler toute l'information apportée par une comparaison systématique de nombreux produits de gènes en un seul graphe général. Cela ne permet toutefois pas d'accéder au résultat précis d'une comparaison. Les résultats complets des comparaisons auxquelles nous avons procédé sont trop volumineux pour être présentés ici dans leur intégralité. Ils sont disponibles à cette adresse : `http://www.bettembourg.fr/PhDocs`. Les tables et figures de cette section montrent par conséquent des extraits des résultats de la comparaison des réactions communes entre ces deux espèces.

2.1 Comparaison entre *Homo sapiens* et *Mus musculus*

Pour chaque réaction commune à *Homo sapiens* et *Mus musculus*, nous avons constitué un ensemble de termes GO par espèce rassemblant toutes les annotations des gènes impliqués. Par exemple, la réaction "*Formation of Malonyl-CoA from Acetyl-CoA (liver)*" peut être catalysée par l'Acetyl-CoA carboxylase 1 (gène ACACA) ou 2 (gène ACACB) chez l'Homme et leurs orthologues chez la souris (gènes Acaca et Acacb). Pour

cette réaction, nous avons constitué un ensemble de termes GO rassemblant toutes les annotations des gènes humains ACACA et ACACB que nous avons comparé à l'ensemble des termes GO rassemblant toutes les annotations des gènes murins Acaca et Acacb.

La Figure 9 résume les éléments nécessaires à obtenir depuis une base de données de voies métaboliques (ici Reactome), une base de données d'identifiants de produits de gènes (ici UniProt) et une base de données d'annotations (ici Gene Ontology Annotation) en vue de mettre en œuvre les mesures de similarité et particularité sémantiques.

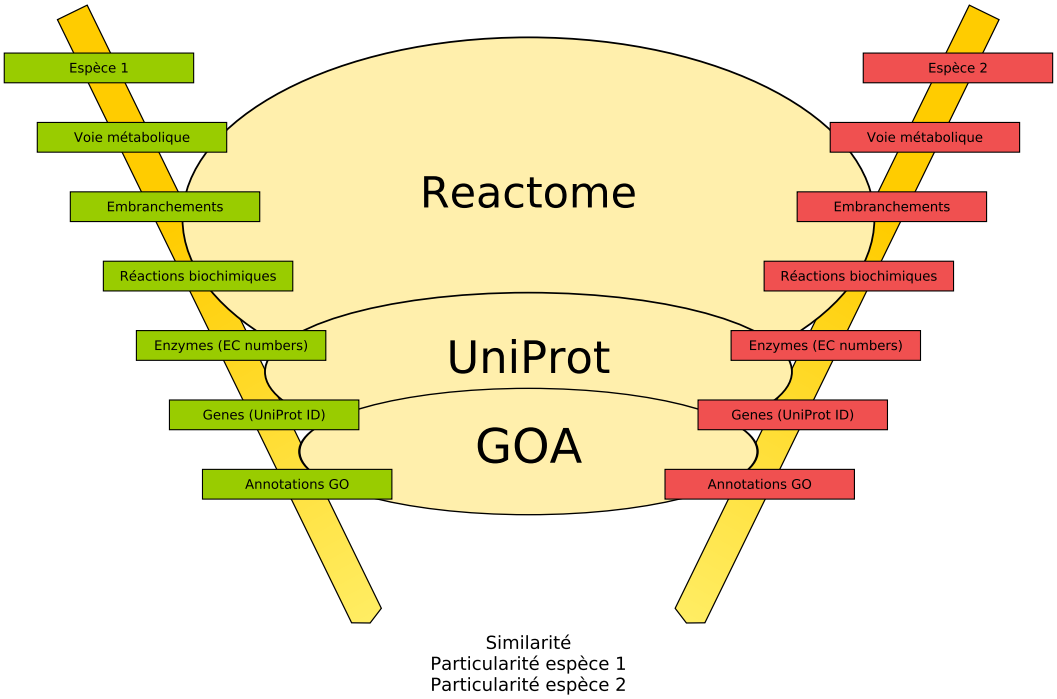


FIGURE 9 – Étapes de la collecte des données en vue de la comparaison des voies métaboliques de deux espèces.

Sur les 650 réactions du métabolisme des lipides communes entre *Homo sapiens* et *Mus musculus*, on compte 289 réactions dans lesquelles interviennent des produits de gènes annotés par des termes GO chez les deux espèces. Ces réactions font intervenir 356 isozymes chez *Homo sapiens* et 311 chez *Mus musculus*.

2.1.1 Vue générale

La comparaison fonctionnelle présentée ici et la comparaison structurelle vue dans la section précédente ne sont pas incompatibles. En effet, les informations des mesures de similarité et de particularité peuvent être figurées sur un graphe de comparaison du métabolisme. Cette représentation a l'avantage de donner une vue générale sur la structure d'une voie métabolique tout en distinguant le niveau de similarité et/ou particularité des produits de gènes intervenant dans ses réactions biochimiques.

Ainsi, la Figure 10 présente le résultat de la comparaison du métabolisme des lipides chez *Homo sapiens* et *Mus musculus*, sur lequel la similarité sémantique des réactions communes est représentée par un gradient du blanc (faible similarité) au bleu (forte similarité). On y retrouve des informations très proches de la comparaison structurelle. Néamnmoins, l'utilisation d'une mesure de similarité qui est a pour résultat des valeurs continues permet d'affiner l'analyse.

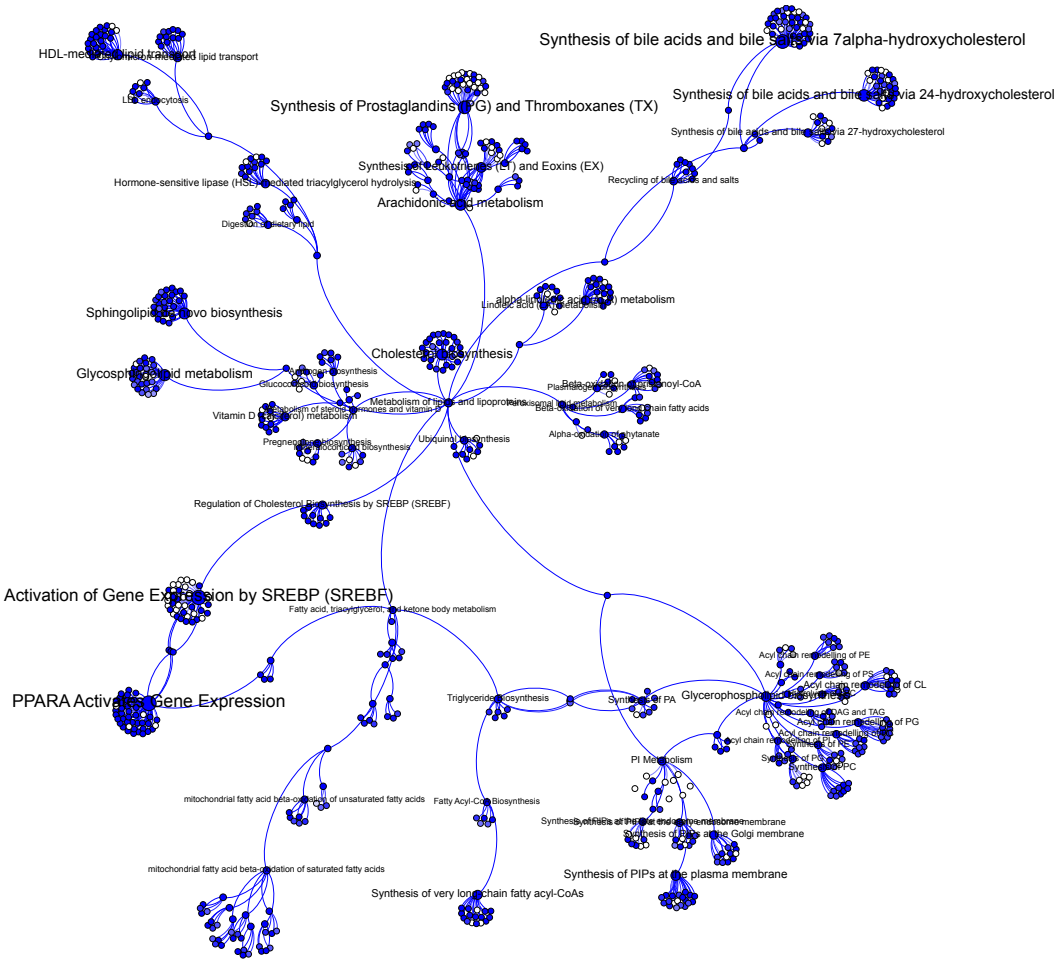


FIGURE 10 – Graphe du métabolisme des lipides de l'Homme et de la souris. La couleur de chaque nœud reflète sa valeur de similarité sémantique. Un nœud est bleu quand il s'agit d'une instance de la classe "*Pathway*" commune aux deux espèces. Plus la similarité entre les deux espèce est forte pour une réaction donnée, plus la feuille correspondante est bleue.

La Figure 11 reprend les parties des Figures 5 et 10 relatives notamment au métabolisme des sphigolipides et des glycolipides. On remarque en particulier que plusieurs réactions qui sont structurellement identiques ont des degrés de similarité qui varient, alors qu'au contraire les réactions relatives à "Sphingolipid de novo biosynthesis" et à "Vitamin D metabolism" sont soit identiques, soit différentes.

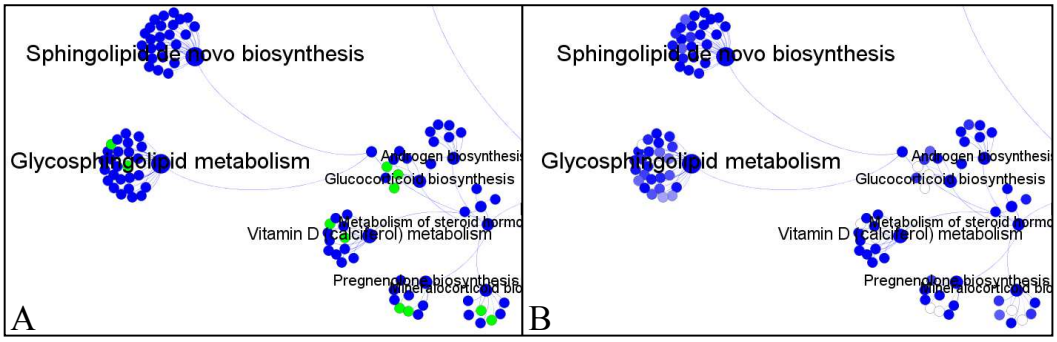


FIGURE 11 – Détail du graphe de la comparaison du métabolisme des lipides de l'Homme et de la souris. La partie A présente un détail de la Figure 5 qui montre la comparaison structurelle entre l'Homme et la souris. La partie B présente un détail de la Figure 10 qui montre la comparaison fonctionnelle basée sur la similarité.

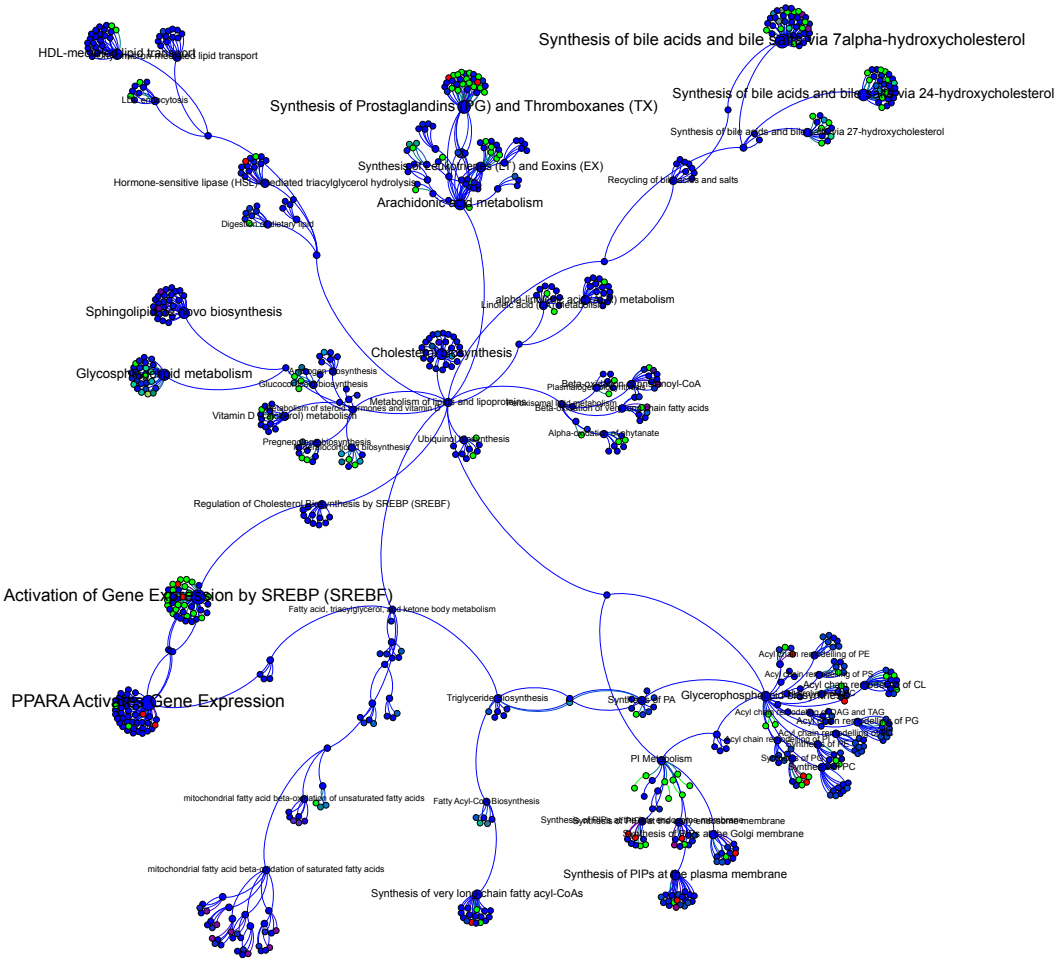


FIGURE 12 – Graphe du métabolisme des lipides de l'Homme et de la souris. La couleur des feuilles reflète la valeur des triplets de similarité et de particularité, dont chaque élément peut varier de façon continue entre 0 et 1. Le bleu reflète une conservation de fonction entre les deux espèces, le vert une particularité humaine et le rouge une particularité murine. Les situations avec une forte similarité associée à une forte particularité chez au moins une des espèces se traduisent donc par des mélanges de couleurs comme bleu-vert ou violet (par exemple dans le quart inférieur gauche).

De la même façon, il est possible de reporter sur le graphe de la comparaison du métabolisme des lipides les informations apportées par nos triplets (similarité, particularité humaine, particularité murine). Ainsi, les feuilles pour lesquelles nous avons obtenu un triplet de valeurs de similarité et de particularités sont colorées en fonction du résultat dans la Figure 12. Pour chaque feuille ainsi colorée, la valeur de similarité est reflétée par la composante bleue, la particularité humaine par la composante verte et la particularité murine par la composante rouge. Plus la couleur d'une feuille tire vers le bleu, plus la réaction qu'elle représente est similaire entre les deux espèces, tandis qu'une dominante verte ou rouge indique respectivement une particularité de l'Homme ou de la souris. Une réaction ayant une similarité et des particularité moyennes sera reflétée par une feuille grise.

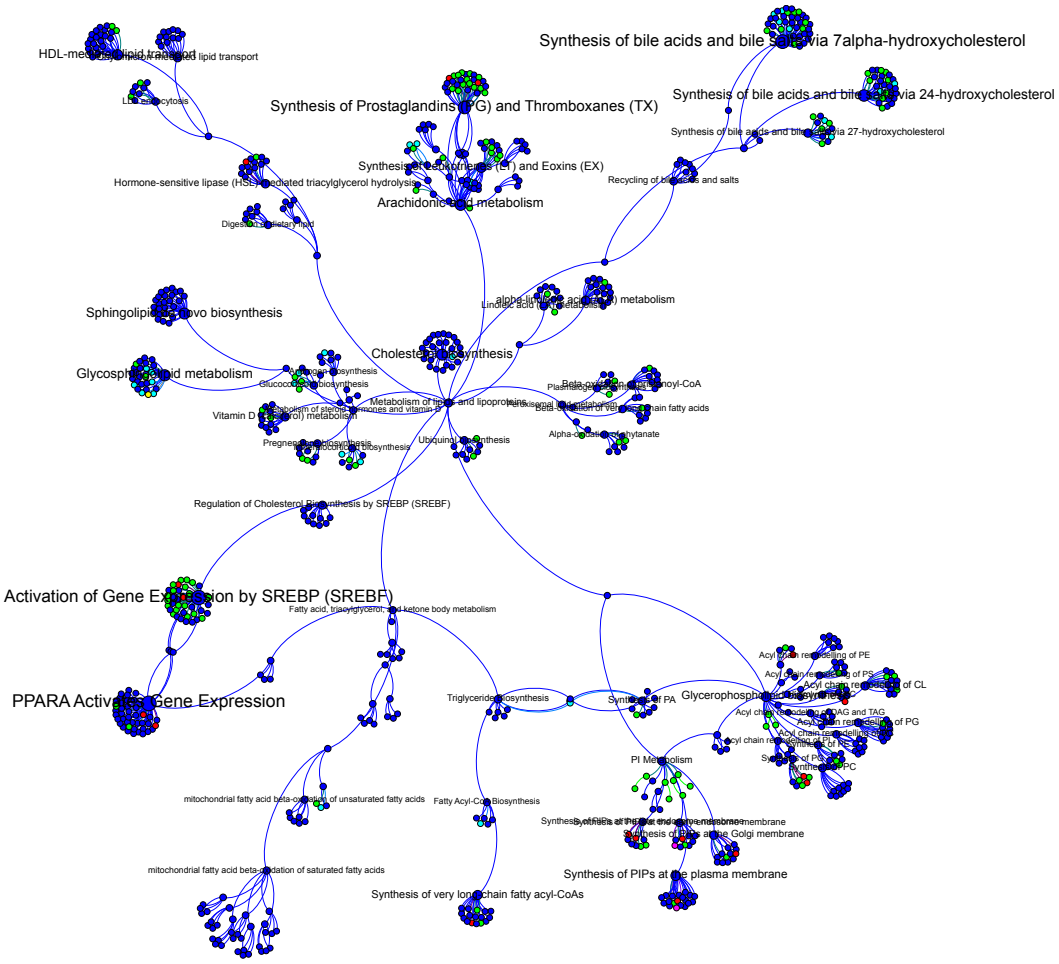


FIGURE 13 – Graphe du métabolisme des lipides de l'Homme et de la souris. La couleur des feuilles reflète le motif résultant de la comparaison. La correspondance motif - couleur est donnée dans la table 5.1.

La Figure 13 discrétise les variations de couleurs des feuilles en appliquant une couleur par motif obtenu avec les triplets de résultat de nos mesures. La similarité est codée par le bleu. Une valeur supérieure au seuil de similarité a pour code RGB `****ff`. La

particularité du premier élément est codée par le vert. Une valeur supérieure au seuil de particularité a pour code RGB `**ff**`. La particularité du second élement est codée par le rouge. Une valeur supérieure au seuil de particularité a pour code RGB : `ff****`. Le motif `+--` a donc pour code RGB `0000ff` (bleu) et le motif `++-` a pour code RGB `00ffff` (bleu-vert). La Table 5.1 fait la correspondance entre les motifs et les couleurs.

Motif	+ + +	+ + -	+ - +	+ - -	- + +	- + -	- - +	- - -
Couleur	Blanc	Bleu-vert	Violet	Bleu	Jaune	Vert	Rouge	Noir

TABLE 5.1 – Correspondance entre les motifs résultant des mesures de similarité et particularité et les couleurs des feuilles dans le graphe des figures 13 et 19.

Les figures 14, 15 et 16 représentent les valeurs de similarité et particularités mesurées respectivement sur BP, MF et CC pour chacune des 289 réactions du métabolisme des lipides dans lesquelles interviennent des produits de gènes annotés par des termes GO chez *Homo sapiens* et *Mus musculus*. Les seuils de similarité (ligne horizontale bleue) et particularité (ligne horizontale jaune) sont reportés sur ces histogrammes, permettant d'identifier directement les différents motifs parmi les résultats. Nous avons identifié chaque réaction par un numéro qui figure en abscisse de ces histogrammes. Grâce à ce type de représentation, il est possible d'avoir une vue d'ensemble des valeurs de similarité et de particularité sur la totalité d'une grande voie métabolique tout en ayant la possibilité d'identifier une réaction dont le résultat de comparaison se démarque par rapport au reste des données.

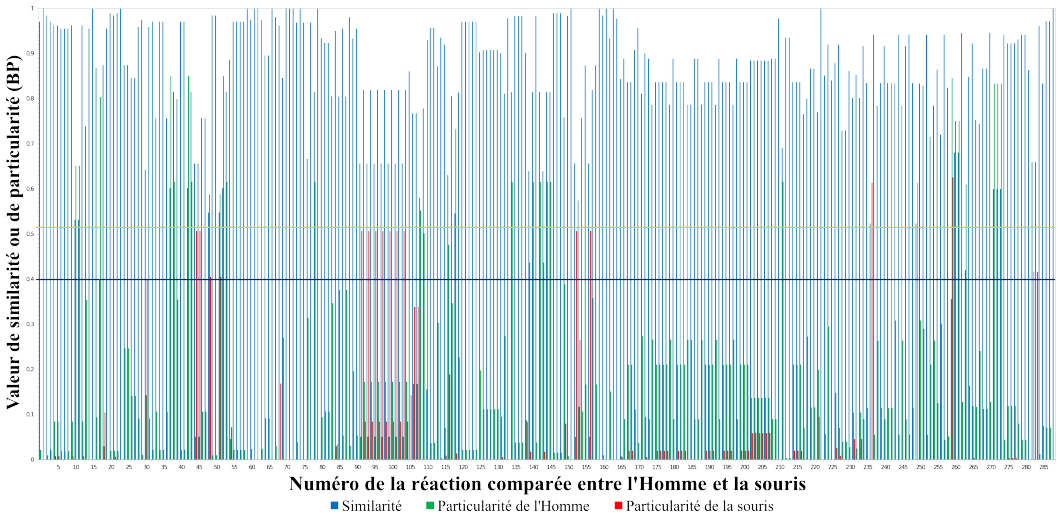


FIGURE 14 – Histogramme des valeurs de similarité et particularités mesurées sur BP pour chacune des 289 réactions du métabolisme des lipides dans lesquelles interviennent des produits de gènes annotés par des termes GO chez *Homo sapiens* et *Mus musculus*.

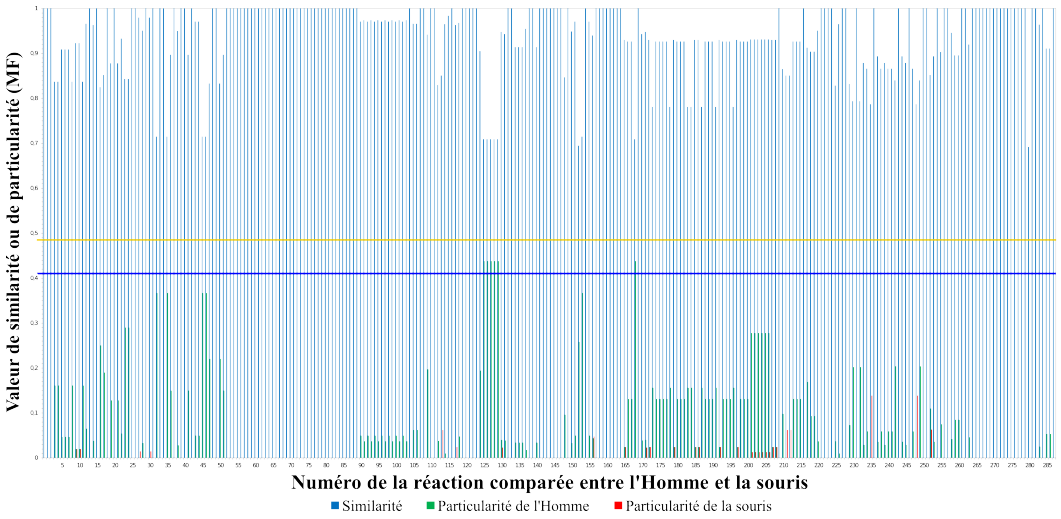


FIGURE 15 – Histogramme des valeurs de similarité et particularités mesurées sur MF pour chacune des 289 réactions du métabolisme des lipides dans lesquelles interviennent des produits de gènes annotés par des termes GO chez *Homo sapiens* et *Mus musculus*.

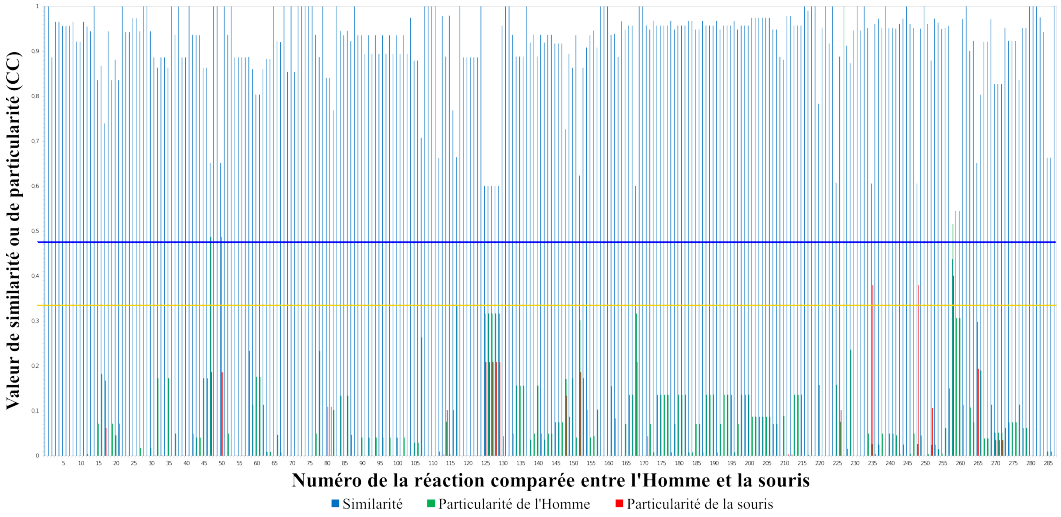


FIGURE 16 – Histogramme des valeurs de similarité et particularités mesurées sur CC pour chacune des 289 réactions du métabolisme des lipides dans lesquelles interviennent des produits de gènes annotés par des termes GO chez *Homo sapiens* et *Mus musculus*.

La table 5.2 fait la synthèse des différents motifs observés dans chaque branche de GO. Les fonctions moléculaires sont particulièrement bien conservées entre *Homo sapiens* et *Mus musculus*. En effet tous les résultats des comparaisons de cette branche se classent comme + - -. Les annotations concernant les composants cellulaires sont également bien conservées entre ces deux espèces. C'est au niveau des processus biologiques qu'on observe le plus de cas de particularité élevée pour l'humain tout en conservant une importante similarité (26). La méthode de comparaison de voies métaboliques

nous permet ainsi d'identifier automatiquement les 30 réactions parmi 289 (soit 10.4 %) présentant des particularités chez l'une ou l'autre des deux espèces sur BP et les 6 réactions (2.1 %) dans le même cas sur CC.

Motif	BP	MF	CC
+ + +	0	0	0
+ + -	26	0	2
+ - +	2	0	2
+ - -	259	289	283
- + +	1	0	1
- + -	1	0	1
- - +	0	0	0
- - -	0	0	0

TABLE 5.2 – Répartition des résultats de la comparaison du métabolisme des lipides chez *Homo sapiens* et *Mus musculus* en motifs en fonction des valeurs de similarité et de particularité.

Les trois cas où la valeur de particularité pour la souris est supérieure au seuil de particularité τ_{par} au niveau de BP sont également ceux où la particularité dépasse τ_{par} au niveau de CC. Il s'agit de deux réactions du métabolisme des phospholipides catalysés par les mêmes enzymes et d'une réaction du métabolisme des sphingolipides. La table 5.3 donne l'extrait des résultats correspondant.

	BP				MF				CC					
Index	Sim	Par hsa	Par mmu	Motif	Sim	Par hsa	Par mmu	Motif	Sim	Par hsa	Par mmu	Motif	Gènes hsa	Gènes mmu
237	0.524	0	0.614	+ - +	0.787	0	0.139	+ - -	0.606	0.026	0.38	+ - +	Q96PE3, O15327	Q6P1Y8, Q9EPW0
250	0.524	0	0.614	+ - +	0.787	0	0.139	+ - -	0.606	0.026	0.38	+ - +	Q96PE3, O15327	Q6P1Y8, Q9EPW0
260	0.356	0.847	0.626	- + +	0.946	0.042	0	+ - -	0.438	0.516	0.401	- + +	P15289	P50428

TABLE 5.3 – Cas de forte particularité de la souris par rapport à l'Homme. L'index n°237 correspond à la réaction"*PI(3,4)P2 is dephosphorylated to PI3P by INPP4A/B at the early endosome membrane*". L'index n°250 correspond à la réaction "*PI(3,4)P2 is dephosphorylated to PI3P by INPP4A/B at the plasma membrane*". L'index n°260 correspond à la réaction "*Arylsulfatase A hydrolyses sulfate from sulfatide to form cerebroside*". Q96PE3 et O15327 sont les Types I et II inositol 3,4-bisphosphate 4-phosphatase de l'Homme (gènes INPP4A et INPP4B) et Q9EPW0 et Q6P1Y8 sont leurs orthologues chez la souris. P15289 est l'Arylsulfatase A de l'Homme (gène ARSA) et P50428 son orthologue murin.

2.1.2 Extrait des résultats

Afin d'avoir un aperçu concret des résultats obtenus, nous avons sélectionné une des branches du métabolisme des lipides. La table 5.4 montre l'organisation du métabolisme de l'acide arachidonique. La colonne de droite contient les réactions biochimiques de cette voie métabolique. Pour chacune de ces réactions, nous avons calculé la similarité et la particularité sémantiques des ensembles de termes GO correspondant à l'annotation des gènes présents pour chaque espèce.

Arachidonic acid metabolism	Synthesis of 12-eicosatetraenoic acid derivatives	12R-HpETE is reduced to 12R-HETE by GPX1/2/4
		12S-HpETE is reduced to 12S-HETE by GPX1/2/4
		Arachidonic acid is converted to 12-oxoETE by ALOX12
		Arachidonic acid is oxidised to 12R-HpETE by ALOX12B
		Arachidonic acid is oxidised to 12S-HpETE by ALOX12/15
		15S-HpETE is reduced to 15S-HETE by GPX1/2/4
		Arachidonic acid is oxidised to 15R-HETE by Acetyl-PTGS2
		Arachidonic acid is oxidised to 15S-HpETE by ALOX15/15B
	Synthesis of 5-eicosatetraenoic acids	5S-HpETE is reduced to 5S-HETE by GPX1/2/4
	Synthesis of epoxy (EET) and dihydroxyeicosatrienoic acids (DHET)	EET(1) is hydrolysed to DHET(1) by EPHX2
	Synthesis of Leukotrienes (LT) and Eoxins (EX)	5S-HpETE is dehydrated to LTA4 by ALOX5
		LTA4 is converted to LTC4 by LTC4S
		LTA4 is hydrolyzed to LTB4
		LTB4 is hydroxylated to 20oh-LTB4 by CYP4F2/4F3
		LTC4 is converted to LTD4 by GGT1/5
		Oxidation of arachidonic acid to 5-HpETE
	Synthesis of Lipoxins (LX)	LXA4 is oxidised to 15k-LXA4 by HPGD
	Synthesis of Prostaglandins (PG) and Thromboxanes (TX)	Arachidonic acid oxidised to PGG2
		PGD2/E2/F2a is oxidised to 15k-PGD2/E2/F2a by HPGD
		PGE2 is converted to PGF2a by CBR1
		PGG2 is reduced to PGH2 by PTGS1
		PGG2 is reduced to PGH2 by PTGS2
		PGH2 is isomerised to PGD2 by HPGDS
		PGH2 is isomerised to PGD2 by PTGDS
		PGH2 is isomerised to PGE2 by PTGES
		PGH2 is isomerised to PGI2 by PTGIS
		PGH2 is isomerised to TXA2 by TBXAS1
		Prostaglandin E synthase isomerizes PGH2 to PGE2

TABLE 5.4 – Organisation du métabolisme de l'acide arachidonique. Les deux premières colonnes contiennent un nom d'embranchement de la voie métabolique de l'acide arachidonique. Chacun de ces embranchement est figuré par un nœud dans le graphe de la Figure 5. La colonne de droite contient les réactions biochimiques qui constituent les étapes de cette voie métabolique. Chacune de ces réactions est figurée par une feuille dans le graphe de la Figure 5.

L'annotation est obtenue à partir de la table Uniprot de GOA via les identifiants Uniprot qui correspondent pour chaque espèce au(x) code(s) EC trouvé(s) pour chaque réaction[Dimmer et al., 2012]. La table 5.5 liste ces identifiants pour le métabolisme de l'acide arachidonique.

Réaction	Code EC	Identifiants Uniprot des gènes hsa correspondants	Identifiants Uniprot des gènes mmu correspondants
12R-HpETE is reduced to 12R-HETE by GPX1/2/4	1.11.1.9	P59796, P18283, Q8TED1, P07203, Q96SL4, P30041, O75715, P22352	O08709, P46412, Q9D7B7, Q9JHC0, P21765, Q99LJ6, P11352, Q91WR8
12S-HpETE is reduced to 12S-HETE by GPX1/2/4	1.11.1.9	P59796, P18283, Q8TED1, P07203, Q96SL4, P30041, O75715, P22352	O08709, P46412, Q9D7B7, Q9JHC0, P21765, Q99LJ6, P11352, Q91WR8
Arachidonic acid is converted to 12-oxoETE by ALOX12	1.13.11.31	P18054, P16050	P39654, P39655, P55249
Arachidonic acid is oxidised to 12R-HpETE by ALOX12B	1.13.11.31	P18054, P16050	P39654, P39655, P55249
Arachidonic acid is oxidised to 12S-HpETE by ALOX12/15	1.13.11.31	P18054, P16050	P39654, P39655, P55249
15S-HpETE is reduced to 15S-HETE by GPX1/2/4	1.11.1.9	P59796, P18283, Q8TED1, P07203, Q96SL4, P30041, O75715, P22352	O08709, P46412, Q9D7B7, Q9JHC0, P21765, Q99LJ6, P11352, Q91WR8
Arachidonic acid is oxidised to 15R-HETE by Acetyl-PTGS2	1.13.11.33	P16050, O15296	P39654
Arachidonic acid is oxidised to 15S-HpETE by ALOX15/15B	1.13.11.33	P16050, O15296	P39654
5S-HpETE is reduced to 5S-HETE by GPX1/2/4	1.11.1.9	P59796, P18283, Q8TED1, P07203, Q96SL4, P30041, O75715, P22352	O08709, P46412, Q9D7B7, Q9JHC0, P21765, Q99LJ6, P11352, Q91WR8
EET(1) is hydrolysed to DHET(1) by EPHX2	3.3.2.10	P34913	P34914
5S-HpETE is dehydrated to LTA4 by ALOX5	1.13.11.34	E5FPY5, E5FPY7, P09917, E5FPY8	P48999
LTA4 is converted to LTC4 by LTC4S	4.4.1.20	Q16873	Q60860
LTA4 is hydrolyzed to LTB4	3.3.2.6	P09960	P24527
LTB4 is hydroxylated to 20oh-LTB4 by CYP4F2/4F3	1.14.13.30	P78329, Q08477	Q99N16, Q9EP75
LTC4 is converted to LTD4 by GGT1/5	2.3.2.2	Q9UJ14, O76032, A6NGU5, P36269, P36268, O75693, Q6P531, P19440	Q6PDE7, Q60928, Q99JP7, Q9Z2A9
Oxidation of arachidonic acid to 5-HpETE	1.13.11.34	E5FPY5, E5FPY7, P09917, E5FPY8	P48999
LXA4 is oxidised to 15k-LXA4 by HPGD	1.1.1.141	P15428	Q8VCC1
Arachidonic acid oxidised to PGG2	1.14.99.1	Q9NNY7, Q6LCE7, P35354, D9MWI3, P23219	P22437, Q05769
PGD2/E2/F2a is oxidised to 15k-PGD2/E2/F2a by HPGD	1.1.1.141	P15428	Q8VCC1
PGE2 is converted to PGF2a by CBR1	1.1.1.189	P16152	P48758
PGG2 is reduced to PGH2 by PTGS1	1.11.1.7	Q16771, P11678, Q92626, P22079, A1KZ92	P49290, Q3UQ28
PGG2 is reduced to PGH2 by PTGS2	1.11.1.7	Q16771, P11678, Q92626, P22079, A1KZ92	P49290, Q3UQ28
PGH2 is isomerised to PGD2 by HPGDS	5.3.99.2	O60760, P41222	O09114, Q9JHF7
PGH2 is isomerised to PGD2 by PTGDS	5.3.99.2	O60760, P41222	O09114, Q9JHF7
PGH2 is isomerised to PGE2 by PTGES	5.3.99.3	Q15185, O14684, Q9H7Z7	Q8BWM0, Q9JM51, Q9R0Q7
PGH2 is isomerised to PGI2 by PTGIS	5.3.99.4	Q6LEN0, Q6LEN2, Q16647	O35074
PGH2 is isomerised to TXA2 by TBXAS1	5.3.99.5	Q16843, P24557	P36423
Prostaglandin E synthase isomerizes PGH2 to PGE2	5.3.99.3	Q15185, O14684, Q9H7Z7	Q8BWM0, Q9JM51, Q9R0Q7

TABLE 5.5 – Enzymes intervenant dans chaque réaction du métabolisme de l'acide arachidonique chez *Homo sapiens* et *Mus musculus*. Certaines enzymes catalysent plusieurs réactions.

La table 5.6 donne les valeurs de similarité et de particularité mesurées entre l'Homme et la souris au niveau de chaque étape du métabolisme de l'acide arachidonique, ainsi que le motif (ou "*pattern*", terme utilisé dans l'article du chapitre 4) correspondant. Le premier symbole du motif concerne la similarité entre les produits de gènes intervenant pour chaque espèce dans la réaction, le deuxième et troisième correspondent aux particularités respectives des produits de gènes de chaque espèce. Dans les réactions "*Arachidonic acid is oxidised to 15R-HETE by Acetyl-PTGS2*", "*Arachidonic acid is oxidised to 15S-HpETE by ALOX15/15B*" et "*LTB4 is hydroxylated to 20oh-LTB4 by CYP4F2/4F3*", on observe un motif + + - au niveau des processus biologiques. Parmi ces réactions, les deux premières sont catalysées par l'enzyme EC 1.13.11.33 (P16050, soit l'Arachidonate 15-lipoxygenase et O15296, soit l'Arachidonate 15-lipoxygenase B chez l'Homme et P39654, soit l'Arachidonate 15-lipoxygenase chez la souris) et la troisième par l'enzyme EC 1.14.13.30 (P78329, soit la Leukotriene-B(4) omega-hydroxylase 1 et Q08477, soit la Leukotriene-B(4) omega-hydroxylase 2 chez l'Homme et Q99N16, soit la Leukotriene-B(4) omega-hydroxylase 2 et Q9EP75, soit la Leukotriene-B4 omega-hydroxylase 3 chez la souris). Les annotations de ces enzymes témoignent donc de leurs implications dans des processus biologiques supplémentaires chez l'Homme comparativement à la souris. Toutes les autres comparaisons sur BP, ainsi que la totalité des comparaisons sur MF et sur CC ont eu pour résultat une forte similarité sans particularité pour l'une ou l'autre des espèces comparées.

	BP				MF				CC			
Réaction	Sim	Par(hsa)	Par(mmu)	Motif BP	Sim	Par(hsa)	Par(mmu)	Motif MF	Sim	Par(hsa)	Par(mmu)	Motif CC
12R-HpETE is reduced to 12R-HETE by GPX1/2/4	0.963	0.085	0.008	+ - -	0.837	0.161	0	+ - -	0.966	0	0	+ - -
12S-HpETE is reduced to 12S-HETE by GPX1/2/4	0.963	0.085	0.008	+ - -	0.837	0.161	0	+ - -	0.966	0	0	+ - -
Arachidonic acid is converted to 12-oxoETE by ALOX12	0.956	0.019	0	+ - -	0.909	0.047	0	+ - -	0.957	0	0	+ - -
Arachidonic acid is oxidised to 12R-HpETE by ALOX12B	0.956	0.019	0	+ - -	0.909	0.047	0	+ - -	0.957	0	0	+ - -
Arachidonic acid is oxidised to 12S-HpETE by ALOX12/15	0.956	0.019	0	+ - -	0.909	0.047	0	+ - -	0.957	0	0	+ - -
15S-HpETE is reduced to 15S-HETE by GPX1/2/4	0.963	0.085	0.008	+ - -	0.837	0.161	0	+ - -	0.966	0	0	+ - -
Arachidonic acid is oxidised to 15R-HETE by Acetyl-PTGS2	0.532	0.652	0	+ + -	0.923	0.02	0.02	+ - -	0.922	0	0	+ - -
Arachidonic acid is oxidised to 15S-HpETE by ALOX15/15B	0.532	0.652	0	+ + -	0.923	0.02	0.02	+ - -	0.922	0	0	+ - -
5S-HpETE is reduced to 5S-HETE by GPX1/2/4	0.963	0.085	0.008	+ - -	0.837	0.161	0	+ - -	0.966	0	0	+ - -
EET(1) is hydrolysed to DHET(1) by EPHX2	0.739	0.354	0	+ - -	0.966	0.065	0	+ - -	0.956	0.004	0	+ - -
5S-HpETE is dehydrated to LTA4 by ALOX5	0.956	0	0	+ - -	1	0	0	+ - -	0.945	0	0	+ - -
LTA4 is converted to LTC4 by LTC4S	1	0	0	+ - -	0.963	0.038	0	+ - -	1	0	0	+ - -
LTA4 is hydrolyzed to LTB4	0.869	0.094	0	+ - -	1	0	0	+ - -	0.836	0.072	0	+ - -
LTB4 is hydroxylated to 20oh-LTB4 by CYP4F2/4F3	0.4	0.804	0	+ + -	0.825	0.25	0	+ - -	0.868	0.182	0	+ - -
LTC4 is converted to LTD4 by GGT1/5	0.875	0.03	0.105	+ - -	0.852	0.19	0	+ - -	0.74	0.168	0.062	+ - -
Oxidation of arachidonic acid to 5-HpETE	0.956	0	0	+ - -	1	0	0	+ - -	0.945	0	0	+ - -
LXA4 is oxidised to 15k-LXA4 by HPGD	0.99	0.02	0	+ - -	0.878	0.128	0	+ - -	0.836	0.072	0	+ - -
Arachidonic acid oxidised to PGG2	0.985	0.02	0.007	+ - -	1	0	0	+ - -	0.882	0.046	0	+ - -
PGD2/E2/F2a is oxidised to 15k-PGD2/E2/F2a by HPGD	0.99	0.02	0	+ - -	0.878	0.128	0	+ - -	0.836	0.072	0	+ - -
PGE2 is converted to PGF2a by CBR1	1	0	0	+ - -	0.933	0.054	0	+ - -	1	0	0	+ - -
PGG2 is reduced to PGH2 by PTGS1	0.875	0.247	0	+ - -	0.843	0.29	0	+ - -	0.943	0	0	+ - -
PGG2 is reduced to PGH2 by PTGS2	0.875	0.247	0	+ - -	0.843	0.29	0	+ - -	0.943	0	0	+ - -
PGH2 is isomerised to PGD2 by HPGDS	0.846	0.142	0	+ - -	1	0	0	+ - -	0.974	0	0	+ - -
PGH2 is isomerised to PGD2 by PTGDS	0.846	0.142	0	+ - -	1	0	0	+ - -	0.974	0	0	+ - -
PGH2 is isomerised to PGE2 by PTGES	0.959	0.091	0	+ - -	0.98	0	0.015	+ - -	0.945	0.018	0	+ - -
PGH2 is isomerised to PGI2 by PTGIS	0.975	0.011	0	+ - -	0.951	0.033	0	+ - -	1	0	0	+ - -
PGH2 is isomerised to TXA2 by TBXAS1	0.643	0.143	0.4	+ - -	1	0	0	+ - -	1	0	0	+ - -
Prostaglandin E synthase isomerizes PGH2 to PGE2	0.959	0.091	0	+ - -	0.98	0	0.015	+ - -	0.945	0.018	0	+ - -

TABLE 5.6 – Valeurs de similarité et de particularité sémantiques mesurées sur BP, MF et CC pour chaque étape du métabolisme de l'acide arachidonique entre *Homo sapiens* et *Mus musculus*. Les motifs correspondants sont indiqués.

2.2 Comparaison entre *Homo sapiens* et *Gallus gallus*

Nous avons comparé le métabolisme des lipides chez *Homo sapiens* et *Gallus gallus* de la même façon, en mesurant la similarité et la particularité sémantiques des ensembles d'annotations des produits de gènes impliqués dans chaque réaction commune aux deux espèces. Sur les 556 réactions du métabolisme des lipides communes entre *Homo sapiens* et *Gallus gallus*, on compte 216 réactions dans lesquelles interviennent des produits de gènes annotés par des termes GO chez les deux espèces. Ces réactions font intervenir 282 isozymes chez *Homo sapiens* et 108 chez *Gallus gallus*.

2.2.1 Vue générale

La Figure 17 présente le résultat de la comparaison du métabolisme des lipides chez *Homo sapiens* et *Gallus gallus*, sur laquelle la similarité sémantique des réactions communes est représentée par un gradient du blanc (faible similarité) au bleu (forte similarité).

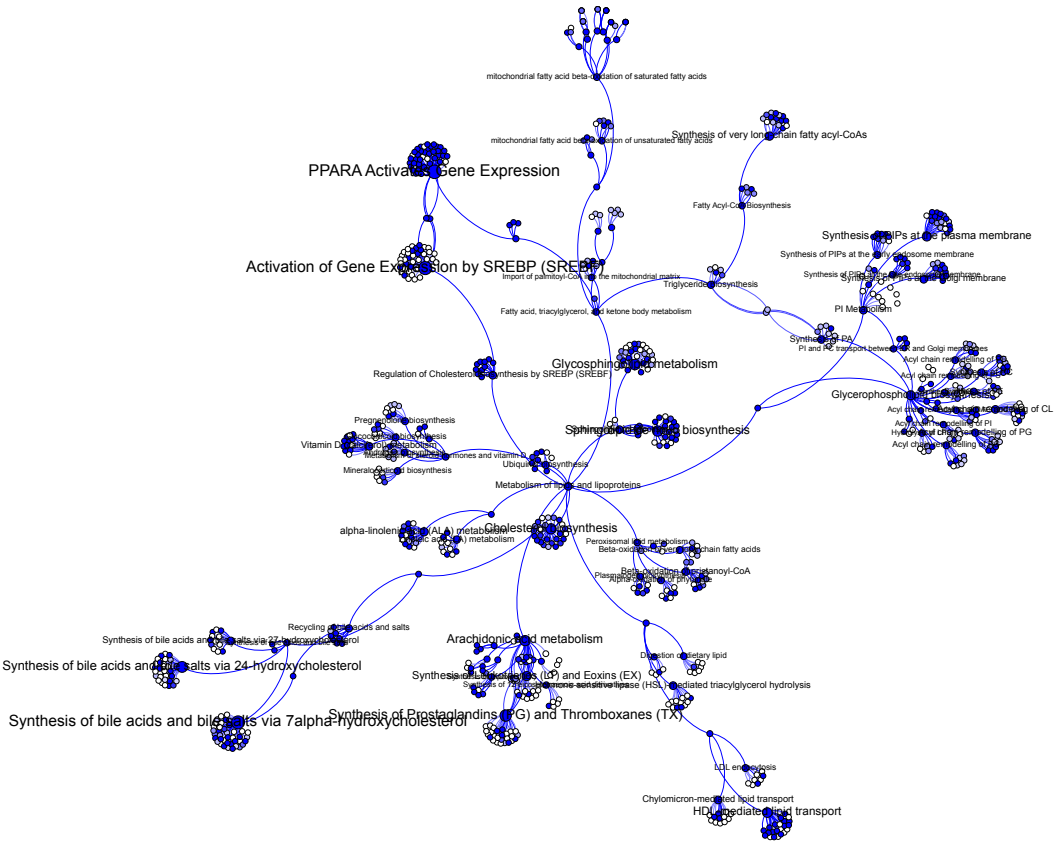


FIGURE 17 – Graphe du métabolisme des lipides de l'Homme et de la poule. La couleur de chaque nœud reflète sa valeur de similarité sémantique. Un nœud est bleu quand il s'agit d'une instance de la classe "*Pathway*" commune aux deux espèces. Plus la similarité entre les deux espèce est forte pour une réaction donnée, plus la feuille correspondante est bleue.

La Figure 18 reprend la comparaison structurelle du métabolisme des lipides entre l'Homme et la poule en colorant les feuilles pour lesquelles nous avons obtenu un triplet de valeurs de similarité et de particularités en fonction de ce triplet. Pour chaque feuille ainsi colorée, la valeur de similarité est reflétée par la composante bleue, la particularité humaine par la composante verte et la particularité de la poule par la composante rouge. Ainsi, plus la couleur d'une feuille tire vers le bleu, plus la réaction qu'elle représente est similaire entre les deux espèces, tandis qu'une dominante verte ou rouge indique respectivement une particularité de l'Homme ou de la poule. Une réaction ayant une similarité et des particularités moyennes sera reflétée par une feuille grise.

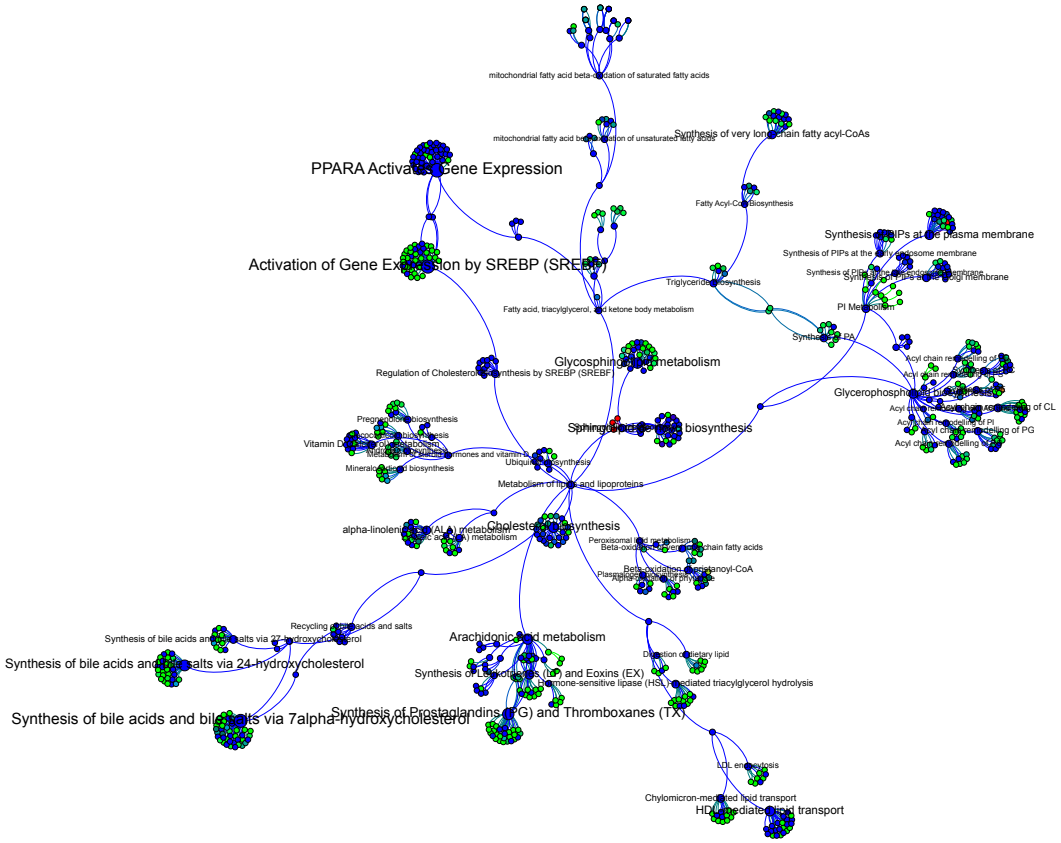


FIGURE 18 – Graphe du métabolisme des lipides de l'Homme et de la poule. La couleur des feuilles reflète la valeur des triplets de similarité et particularités. Le bleu reflète une conservation de fonction entre les deux espèces, le vert une particularité humaine et le rouge une particularité de la poule.

La Figure 19 discrétise les variations de couleurs des feuilles en appliquant une couleur par motif obtenu avec les triplets résultant de nos mesures. La Table 5.1 fait la correspondance entre les motifs et les couleurs.

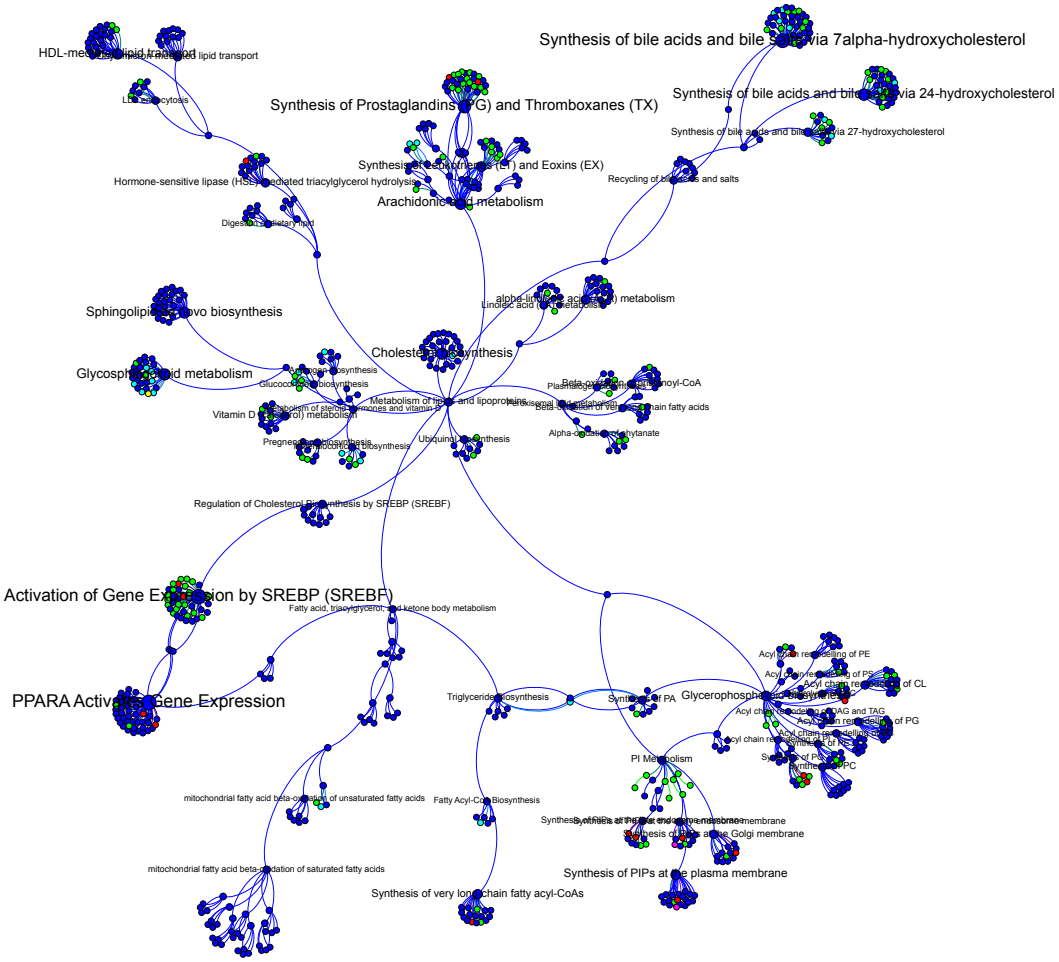


FIGURE 19 – Graphe du métabolisme des lipides de l'Homme et de la poule. La couleur des feuilles reflète le motif résultant de la comparaison. La correspondance motif - couleur est donnée dans la table 5.1.

Les figures 20, 21 et 22 représentent les valeurs de similarité et particularités mesurées respectivement sur BP, MF et CC pour chacune des 216 réactions du métabolisme des lipides dans lesquelles interviennent des produits de gènes annotés par des termes GO chez *Homo sapiens* et *Gallus gallus*. Les seuils de similarité (ligne horizontale bleue) et particularité (ligne horizontale jaune) sont reportés sur ces histogrammes, permettant d'identifier directement les différents motifs parmi les résultats. Nous avons identifié chaque réaction par un numéro qui figure en abscisse de ces histogrammes.

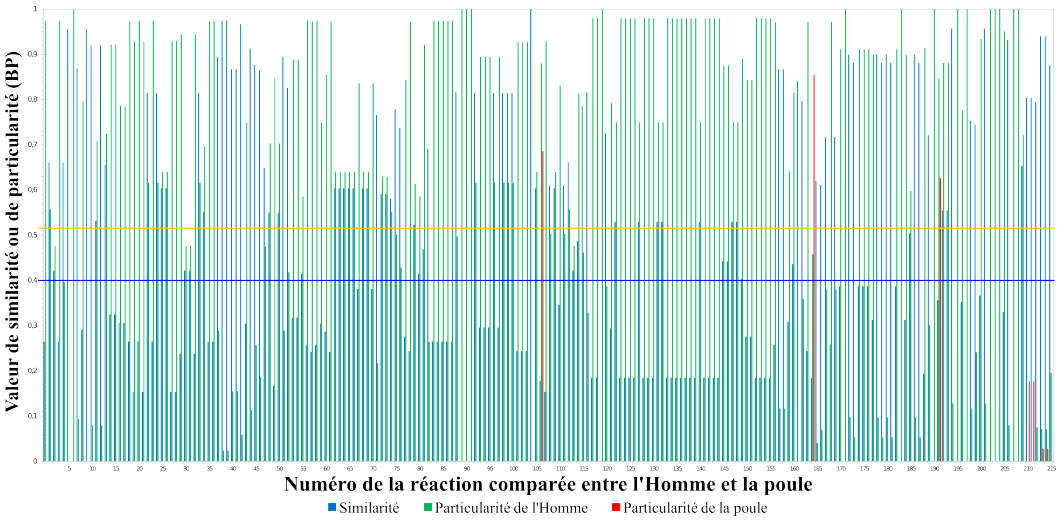


FIGURE 20 – Histogramme des valeurs de similarité et particularités mesurées sur BP pour le métabolisme des lipides chez l'Homme et la poule.

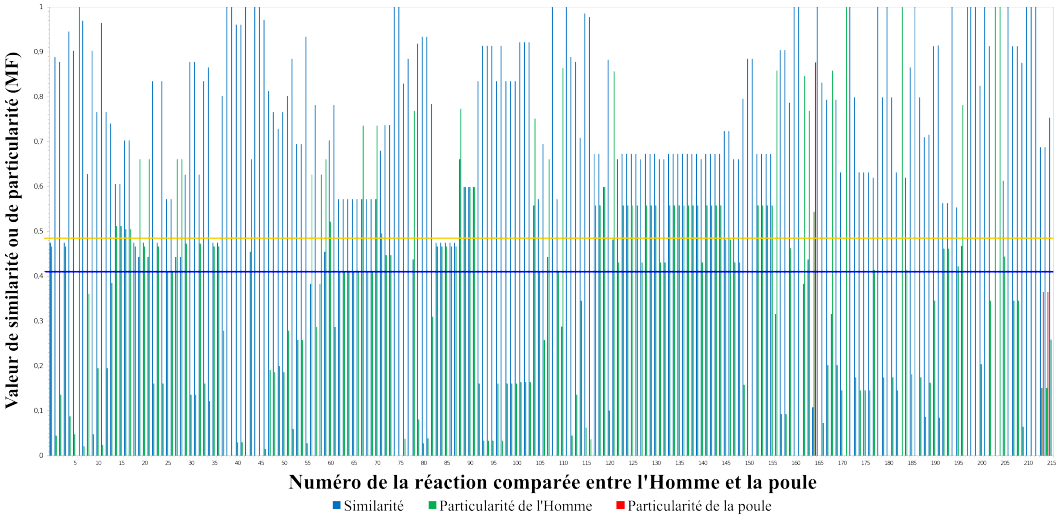


FIGURE 21 – Histogramme des valeurs de similarité et particularités mesurées sur MF pour le métabolisme des lipides chez l'Homme et la poule.

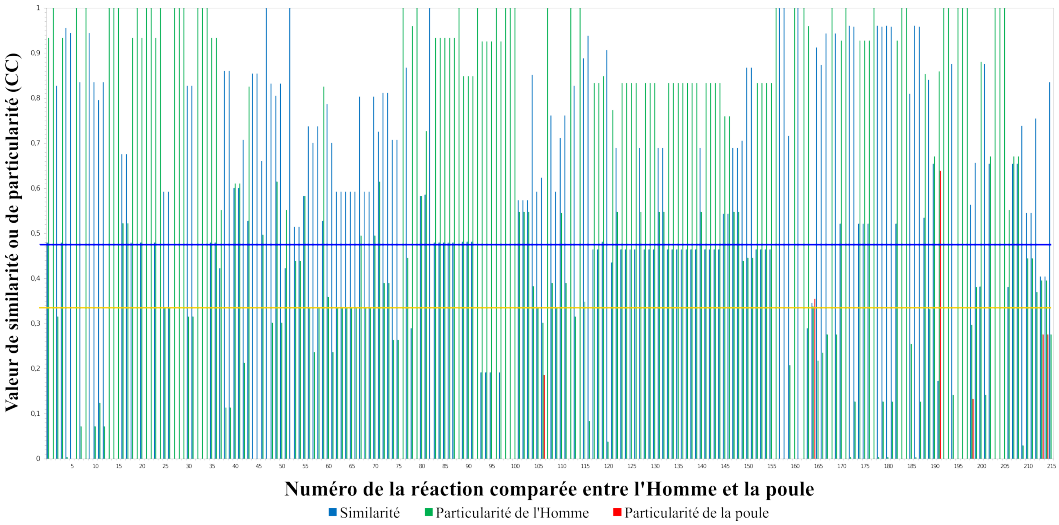


FIGURE 22 – Histogramme des valeurs de similarité et particularités mesurées sur CC pour le métabolisme des lipides chez l'Homme et la poule.

La table 5.7 fait la synthèse des différents motifs observés dans chaque branche de GO. Comme pour la comparaison entre l'Homme et la souris, les annotations relatives aux fonctions moléculaires sont les mieux conservées entre *Homo sapiens* et *Gallus gallus*. Par contre, alors que les valeurs de similarités en dessous du seuil étaient rares lors de la comparaison précédente, elles sont très fréquentes ici sur BP et CC, majoritairement dans des motifs - + -. Seuls six cas de valeur de particularité supérieure au seuil pour la poule ont été mesurés (trois sur BP, un sur MF et deux sur CC).

Motif	BP	MF	CC
+ + +	0	0	0
+ + -	49	49	83
+ - +	0	0	0
+ - -	61	154	51
- + +	2	1	2
- + -	103	12	80
- - +	1	0	0
- - -	0	0	0

TABLE 5.7 – Répartition des résultats de la comparaison du métabolisme des lipides chez *Homo sapiens* et *Gallus gallus* en motifs en fonction des valeurs de similarité et de particularité.

Les cas où la valeur de particularité pour la poule est supérieure au seuil de particularité τ_{par} concernent une réaction du métabolisme lipidique du peroxysome, une réaction du métabolisme des phospholipides et une réaction du métabolisme des sphingolipides. La table 5.8 donne l'extrait des résultats correspondant.

	BP				MF				CC					
Index	Sim	Par hsa	Par mmu	Motif	Sim	Par hsa	Par mmu	Motif	Sim	Par hsa	Par mmu	Motif	Gènes hsa	Gènes mmu
106	0.178	0.88	0.686	- + +	0.695	0.258	0	+ - -	0.624	0.302	0.186	+ - -	Q9UKG9	E1BRU9
164	0.184	0.458	0.854	- - +	0.108	0.544	0.877	- + +	0.346	0.338	0.355	- + +	Q99447	F1NC39
191	0.356	0.847	0.626	- + +	0.915	0.085	0	+ - -	0.173	0.86	0.639	- + +	P15289	F1NWF7

TABLE 5.8 – Cas de forte particularité de la poule par rapport à l'Homme. L'index n°106 correspond à la réaction"*4,8-dimethylnonanoyl-CoA + carnitine → 4,8-dimethylnonanoylcarnitine + CoASH*". L'index n°164 correspond à la réaction "*PETA and CTP are condensed to CDP-ETA by PCY2*". L'index n°191 correspond à la réaction "*Arylsulfatase A hydrolyses sulfate from sulfatide to form cerebroside*". Q9UKG9 est la Peroxisomal carnitine O-octanoyltransferase de l'Homme (gène CROT). Q99447 est la Ethanolamine-phosphate cytidylyltransferase de l'Homme (gène PCYT2). P15289 est l'Arylsulfatase A de l'Homme (gène ARSA). Les trois identifiants E1BRU9, F1NC39 et F1NWF7 correspondent chacun à une "*Uncharacterized protein*". E1BRU9 présente des similarités de séquences avec la famille des carnitine/choline acetyltransferases. Le symbole du gène codant pour F1NC39 est PCYT2, le même que celui codant pour le gène humain Q99447 impliqué dans la même réaction. Le symbole du gène codant pour F1NWF7 est ARSA, le même que celui codant pour le gène humain P15289 impliqué dans la même réaction.

2.2.2 Extrait des résultats

Afin d'avoir un aperçu concret des résultats obtenus, nous avons sélectionné une des branches du métabolisme des lipides. La table 5.9 montre l'organisation du métabolisme des acides gras, des triglycérides et des corps cétoniques. Cette table ne contient que les réations pour lesquelles une enzyme annotée chez les deux espèces a été trouvée. La colonne de droite contient les réactions biochimiques de cette voie métabolique. Pour chacune de ces réactions, nous avons calculé la similarité et la particularité sémantiques des ensembles de termes GO correspondant à l'annotation des gènes présents pour chaque espèce.

L'annotation est obtenue à partir de la table Uniprot de GOA via les identifiants Uniprot qui correspondent pour chaque espèce au(x) code(s) EC trouvé(s) pour chaque réaction[Dimmer et al., 2012]. La table 5.10 liste ces identifiants pour le métabolisme des acides gras, des triglycérides et des corps cétoniques.

La table 5.11 donne les valeurs de similarité et de particularité mesurées entre l'Homme et la poule au niveau de chaque étape du métabolisme des acides gras, des triglycérides et des corps cétoniques, ainsi que le motif correspondant. Aucune particularité pour *Gallus gallus* n'a pu être mise en évidence dans cet extrait des résultats de la comparaison du métabolisme des lipides entre l'Homme et la poule quelle que soit la branche de GO considérée. Les annotations de la branche "Molecular functions" sont les mieux conservées entre les deux espèces. Sur les 36 comparaisons de la table 5.11, 28 forment un motif + - - pour MF, contre seulement 6 pour CC et 5 pour BP. Les motifs les plus fréquents sur BP sont - + - (17) et + + - (14). Le motif le plus fréquent sur CC est + + - (27).

Fatty acid, triacylglycerol, and ketone body metabolism				Gly-3-P+FAD->DHAP+FADH2 (catalyzed by mitochondrial Gly-Phos dehydrogenase)
	Import of palmitoyl-CoA into the mitochondrial matrix			CPT1 converts palmitoyl-CoA to palmitoyl carnitine
				CPT2 converts palmitoyl carnitine to palmitoyl-CoA
				Formation of Malonyl-CoA from Acetyl-CoA (muscle)
	Ketone body metabolism	Ketone body catabolism		acetoacetyl-CoA + CoA <=> 2 acetyl-CoA
				D-beta hydroxybutyrate+NAD+ <=> acetoacetate+NADH+H+
		Synthesis of Ketone Bodies		2 acetyl-CoA <=> acetoacetyl-CoA + CoA
				acetoacetyl-CoA+acetyl-CoA => HMG-CoA + CoASH
				HMG CoA => acetoacetic acid + acetyl CoA
				Reduction of Acetoacetate to beta-Hydroxybutyrate
	Mitochondrial Fatty Acid Beta-Oxidation	mitochondrial fatty acid beta-oxidation of saturated fatty acids	Beta oxidation of butanoyl-CoA to acetyl-CoA	(S)-Hydroxybutanoyl-CoA+NAD<=>Acetoacetyl-CoA+NADH+H
			Beta oxidation of decanoyl-CoA to octanoyl-CoA-CoA	(S)-Hydroxydecanoyl-CoA+NAD<=>3-Oxodecanoyl-CoA+NADH+H
			Beta oxidation of hexanoyl-CoA to butanoyl-CoA	(S)-Hydroxyhexanoyl-CoA+NAD<=>3-Oxohexanoyl-CoA+NADH+H
			Beta oxidation of lauroyl-CoA to decanoyl-CoA-CoA	(S)-3-Hydroxydodecanoyl-CoA+NAD<=>3-Oxododecanoyl-CoA+NADH+H
			Beta oxidation of myristoyl-CoA to lauroyl-CoA	(S)-3-Hydroxytetradecanoyl-CoA+NAD<=>3-Oxotetradecanoyl-CoA+NADH+H
				trans-Tetradec-2-enoyl-CoA+H2O<=>(S)-3-Hydroxytetradecanoyl-CoA
			Beta oxidation of octanoyl-CoA to hexanoyl-CoA	(S)-Hydroxyoctanoyl-CoA+NAD<=>3-Oxooctanoyl-CoA+NADH+H
			Beta oxidation of palmitoyl-CoA to myristoyl-CoA	(S)-3-Hydroxyhexadecanoyl-CoA+NAD<=>3-Oxopalmitoyl-CoA+NADH+H
				trans-Hexadec-2-enoyl-CoA+H2O<=>(S)-3-Hydroxyhexadecanoyl-CoA
		mitochondrial fatty acid beta-oxidation of unsaturated fatty acids		Isomerization of cis,cis-3,6-Dodecadienoyl-CoA to form trans,cis-Lauro-2,6-dienoyl-CoA
				Removal of 2 Carbon atoms from trans,cis-Lauro-2,6-dienoyl-CoA to form 4-cis-decenoyl-CoA
				Removal of six carbons from Linoleoyl-CoA to form cis,cis-3,6- Dodecadienoyl-CoA
		Propionyl-CoA catabolism		D-methylmalonyl-CoA <=> L-methylmalonyl-CoA
				MUT isomerises L-MM-CoA to SUCC-CoA
	Triglyceride Biosynthesis			Conversion of Glycerol to Glycerol-3-phosphate
				Conversion of glycerol-3-phosphate to lysophosphatidic acid
				Conversion of Phosphatidic Acid to Diacylglycerol
		Fatty Acyl-CoA Biosynthesis		Conversion of malonyl-CoA and acetyl-CoA to palmitate
				Formation of Malonyl-CoA from Acetyl-CoA (liver)
				Generation of Cytoplasmic Acetyl CoA from Citrate
			Synthesis of very long-chain fatty acyl-CoAs	3-oxooctadecanoyl-CoA (3-oxostearoyl-CoA) + NADPH + H+ => 3-hydroxyoctadecanoyl-CoA + NADP+
				arachidonate + CoASH + ATP => arachidonoyl-CoA + AMP + pyrophosphate + H2O [ACSL4]
				Conversion of palmitic acid to palmitoyl-CoA
				palmitate + CoASH + ATP => palmitoyl-CoA + AMP + pyrophosphate + H2O [ACSL3]
				palmitate + CoASH + ATP => palmitoyl-CoA + AMP + pyrophosphate + H2O [ACSL5]
				palmitate + CoASH + ATP => palmitoyl-CoA + AMP + pyrophosphate + H2O [ACSL6]

TABLE 5.9 – Organisation du métabolisme des acides gras, des triglycérides et des corps cétoniques. Les deux premières colonnes contiennent un nom d'embranchement de cette voie métabolique. Chacun de ces embranchement est figuré par un nœud dans le graphe de la Figure 6. La colonne de droite contient les réactions biochimiques qui constituent les étapes de cette voie métabolique. Chacune de ces réactions est figurée par une feuille dans le graphe de la Figure 6.

Réaction	Codes EC	Identifiants Uniprot des gènes hsa correspondants	Identifiants Uniprot des gènes gga correspondants
Gly-3-P+FAD->DHAP+FADH2 (catalyzed by mitochondrial Gly-Phos dehydrogenase)	1.1.99.5, 1.1.5.3	P43304	F1NCA2
CPT1 converts palmitoyl-CoA to palmitoyl carnitine	2.3.1.21	P23786, Q8TCG5, P50416, Q92523	Q2PP41, Q6B842
CPT2 converts palmitoyl carnitine to palmitoyl-CoA	2.3.1.21	P23786, Q8TCG5, P50416, Q92523	Q2PP41, Q6B842
Formation of Malonyl-CoA from Acetyl-CoA (muscle)	6.4.1.2	Q13085, O00763	Q5MB12, P11029
acetoacetyl-CoA + CoA <=> 2 acetyl-CoA	2.3.1.9	Q9BWD1, P24752	E1C0Q5
D-beta hydroxybutyrate+NAD+ <=> acetoacetate+NADH+H+	1.1.1.30	Q02338, Q9BUT1	Q5ZJZ5
2 acetyl-CoA <=> acetoacetyl-CoA + CoA	2.3.1.9	Q9BWD1, P24752	E1C0Q5
acetoacetyl-CoA+acetyl-CoA => HMG-CoA + CoASH	2.3.3.10	P54868, Q01581	P23228
HMG CoA => acetoacetic acid + acetyl CoA	4.1.3.4	P35914, Q8TB92	P35915
Reduction of Acetoacetate to beta-Hydroxybutyrate	1.1.1.30	Q02338, Q9BUT1	Q5ZJZ5
(S)-Hydroxybutanoyl-CoA+NAD<=>Acetoacetyl-CoA+NADH+H	1.1.1.35	Q16836, Q08426, Q99714	E1BZH9
(S)-Hydroxydecanoyl-CoA+NAD<=>3-Oxodecanoyl-CoA+NADH+H	1.1.1.35	Q16836, Q08426, Q99714	E1BZH9
(S)-Hydroxyhexanoyl-CoA+NAD<=>3-Oxohexanoyl-CoA+NADH+H	1.1.1.35	Q16836, Q08426, Q99714	E1BZH9
(S)-3-Hydroxydodecanoyl-CoA+NAD<=>3-Oxododecanoyl-CoA+NADH+H	1.1.1.35	Q16836, Q08426, Q99714	E1BZH9
(S)-3-Hydroxytetradecanoyl-CoA+NAD<=>3-Oxotetradecanoyl-CoA+NADH+H	1.1.1.35	Q16836, Q08426, Q99714	E1BZH9
trans-Tetradec-2-enoyl-CoA+H2O<=>(S)-3-Hydroxytetradecanoyl-CoA	4.2.1.17	P30084, P40939, Q08426	E1C1T9
(S)-Hydroxyoctanoyl-CoA+NAD<=>3-Oxooctanoyl-CoA+NADH+H	1.1.1.35	Q16836, Q08426, Q99714	E1BZH9
(S)-3-Hydroxyhexadecanoyl-CoA+NAD<=>3-Oxopalmitoyl-CoA+NADH+H	1.1.1.35	Q16836, Q08426, Q99714	E1BZH9
trans-Hexadec-2-enoyl-CoA+H2O<=>(S)-3-Hydroxyhexadecanoyl-CoA	4.2.1.17	P30084, P40939, Q08426	E1C1T9
isomerization of cis,cis-3,6-Dodecadienoyl-CoA to form trans,cis-Lauro-2,6-dienoyl-CoA	5.3.3.8	P42126, O75521, Q08426	E1BW06
Removal of 2 Carbon atoms from trans,cis-Lauro-2,6-dienoyl-CoA to form 4-cis-decenoyl-CoA	2.3.1.155, 1.3.99.3, 1.1.1.35, 4.2.1.17	Q16836, P40939, Q08426, P30084, Q99714	E1C1T9, E1BZH9
Removal of six carbons from Linoleoyl-CoA to form cis,cis-3,6- Dodecadienoyl-CoA	2.3.1.155, 1.3.99.3, 1.1.1.35, 4.2.1.17	Q16836, P40939, Q08426, P30084, Q99714	E1C1T9, E1BZH9
D-methylmalonyl-CoA <=> L-methylmalonyl-CoA	5.1.99.1	Q96PE7	E1BSK7
MUT isomerises L-MM-CoA to SUCC-CoA	5.4.99.2	P22033	E1BY22
Conversion of Glycerol to Glycerol-3-phosphate	2.7.1.30	Q14410, Q6ZS86, Q14409, P32189	Q5ZMJ4
Conversion of glycerol-3-phosphate to lysophosphatidic acid	2.3.1.15	Q86UL3, Q9HCL2, Q6NUI2, Q53EU6	Q5ZLL8
Conversion of Phosphatidic Acid to Diacylglycerol	3.1.3.4	Q5VZY2, Q96GM1, Q6T4P5, Q7Z2D5, Q9BQK8, O43688, Q92539, O14495, Q8NEB5, Q14693, O14494	F1NLF4
Conversion of malonyl-CoA and acetyl-CoA to palmitate	2.3.1.85	P49327	P12276
Formation of Malonyl-CoA from Acetyl-CoA (liver)	6.4.1.2	Q13085, O00763	Q5MB12, P11029
Generation of Cytoplasmic Acetyl CoA from Citrate	2.3.3.8	P53396	F1P269
3-oxooctadecanoyl-CoA (3-oxostearoyl-CoA) + NADPH + H+ => 3-hydroxyoctadecanoyl-CoA + NADP+	1.1.1.211	P40939	F1NI29
arachidonate + CoASH + ATP => arachidonoyl-CoA + AMP + pyrophosphate + H2O [ACSL4]	6.2.1.3	O95573, Q5FVE4, O14975, O60488, A6GV77, Q9UKU0, Q96GR2, P33121, Q9ULC5	Q5ZKR7
Conversion of palmitic acid to palmitoyl-CoA	6.2.1.3	O95573, Q5FVE4, O14975, O60488, A6GV77, Q9UKU0, Q96GR2, P33121, Q9ULC5	Q5ZKR7
palmitate + CoASH + ATP => palmitoyl-CoA + AMP + pyrophosphate + H2O [ACSL3]	6.2.1.3	O95573, Q5FVE4, O14975, O60488, A6GV77, Q9UKU0, Q96GR2, P33121, Q9ULC5	Q5ZKR7
palmitate + CoASH + ATP => palmitoyl-CoA + AMP + pyrophosphate + H2O [ACSL5]	6.2.1.3	O95573, Q5FVE4, O14975, O60488, A6GV77, Q9UKU0, Q96GR2, P33121, Q9ULC5	Q5ZKR7
palmitate + CoASH + ATP => palmitoyl-CoA + AMP + pyrophosphate + H2O [ACSL6]	6.2.1.3	O95573, Q5FVE4, O14975, O60488, A6GV77, Q9UKU0, Q96GR2, P33121, Q9ULC5	Q5ZKR7

TABLE 5.10 – Enzymes intervenant dans chaque réaction du métabolisme des acides gras, des triglycérides et des corps cétoniques chez *Homo sapiens* et *Gallus gallus*. Certaines enzymes catalysent plusieurs réactions.

	BP				MF				CC			
Réaction	Sim	Par(hsa)	Par(gga)	Motif BP	Sim	Par(hsa)	Par(gga)	Motif MF	Sim	Par(hsa)	Par(gga)	Motif CC
Gly-3-P+FAD->DHAP+FADH2 (catalyzed by mitochondrial Gly-Phos dehydrogenase)	0.826	0.418	0	+ - -	0.885	0.06	0	+ - -	1	0	0	+ - -
CPT1 converts palmitoyl-CoA to palmitoyl carnitine	0.318	0.888	0	- + -	0.695	0.258	0	+ - -	0.515	0.439	0	+ + -
CPT2 converts palmitoyl carnitine to palmitoyl-CoA	0.318	0.888	0	- + -	0.695	0.258	0	+ - -	0.515	0.439	0	+ + -
Formation of Malonyl-CoA from Acetyl-CoA (muscle)	0.415	0.586	0	+ + -	0.934	0.028	0	+ - -	0.583	0.583	0	+ + -
acetoacetyl-CoA + CoA <=> 2 acetyl-CoA	0.257	0.975	0	- + -	0.383	0.627	0	- + -	0.738	0.336	0	+ + -
D-beta hydroxybutyrate+NAD+ <=> acetoacetate+NADH+H+	0.242	0.973	0	- + -	0.782	0.287	0	+ - -	0.701	0.237	0	+ - -
2 acetyl-CoA <=> acetoacetyl-CoA + CoA	0.257	0.975	0	- + -	0.383	0.627	0	- + -	0.738	0.336	0	+ + -
acetoacetyl-CoA+acetyl-CoA => HMG-CoA + CoASH	0.305	0.749	0	- + -	0.455	0.661	0	+ + -	0.528	0.826	0	+ + -
HMG CoA => acetoacetic acid + acetyl CoA	0.287	0.855	0	- + -	0.703	0.522	0	+ + -	0.787	0.359	0	+ + -
Reduction of Acetoacetate to beta-Hydroxybutyrate	0.242	0.973	0	- + -	0.782	0.287	0	+ - -	0.701	0.237	0	+ - -
(S)-Hydroxybutanoyl-CoA+NAD<=>Acetoacetyl-CoA+NADH+H	0.604	0.64	0	+ + -	0.572	0.409	0	+ - -	0.593	0.336	0	+ + -
(S)-Hydroxydecanoyl-CoA+NAD<=>3-Oxodecanoyl-CoA+NADH+H	0.604	0.64	0	+ + -	0.572	0.409	0	+ - -	0.593	0.336	0	+ + -
(S)-Hydroxyhexanoyl-CoA+NAD<=>3-Oxohexanoyl-CoA+NADH+H	0.604	0.64	0	+ + -	0.572	0.409	0	+ - -	0.593	0.336	0	+ + -
(S)-3-Hydroxydodecanoyl-CoA+NAD<=>3-Oxododecanoyl-CoA+NADH+H	0.604	0.64	0	+ + -	0.572	0.409	0	+ - -	0.593	0.336	0	+ + -
(S)-3-Hydroxytetradecanoyl-CoA+NAD<=>3-Oxotetradecanoyl-CoA+NADH+H	0.604	0.64	0	+ + -	0.572	0.409	0	+ - -	0.593	0.336	0	+ + -
trans-Tetradec-2-enoyl-CoA+H2O<=>(S)-3-Hydroxytetradecanoyl-CoA	0.381	0.836	0	- + -	0.572	0.736	0	+ + -	0.804	0.495	0	+ + -
(S)-Hydroxyoctanoyl-CoA+NAD<=>3-Oxooctanoyl-CoA+NADH+H	0.604	0.64	0	+ + -	0.572	0.409	0	+ - -	0.593	0.336	0	+ + -
(S)-3-Hydroxyhexadecanoyl-CoA+NAD<=>3-Oxopalmitoyl-CoA+NADH+H	0.604	0.64	0	+ + -	0.572	0.409	0	+ - -	0.593	0.336	0	+ + -
trans-Hexadec-2-enoyl-CoA+H2O<=>(S)-3-Hydroxyhexadecanoyl-CoA	0.381	0.836	0	- + -	0.572	0.736	0	+ + -	0.804	0.495	0	+ + -
isomerization of cis,cis-3,6-Dodecadienoyl-CoA to form trans,cis-Lauro-2,6-dienoyl-CoA	0.766	0.217	0	+ - -	0.68	0.496	0	+ + -	0.726	0.615	0	+ + -
Removal of 2 Carbon atoms from trans,cis-Lauro-2,6-dienoyl-CoA to form 4-cis-decenoyl-CoA	0.591	0.63	0	+ + -	0.737	0.447	0	+ - -	0.812	0.39	0	+ + -
Removal of six carbons from Linoleoyl-CoA to form cis,cis-3,6- Dodecadienoyl-CoA	0.591	0.63	0	+ + -	0.737	0.447	0	+ - -	0.812	0.39	0	+ + -
D-methylmalonyl-CoA <=> L-methylmalonyl-CoA	0.581	0.552	0	+ + -	1	0	0	+ - -	0.708	0.264	0	+ - -
MUT isomerises L-MM-CoA to SUCC-CoA	0.779	0.502	0	+ - -	1	0	0	+ - -	0.708	0.264	0	+ - -
Conversion of Glycerol to Glycerol-3-phosphate	0.738	0.429	0	+ - -	0.83	0.038	0	+ - -	0	1	0	- + -
Conversion of glycerol-3-phosphate to lysophosphatidic acid	0.276	0.843	0	- + -	0.885	0	0	+ - -	0.868	0.446	0	+ + -
Conversion of Phosphatidic Acid to Diacylglycerol	0.244	0.972	0	- + -	0.438	0.769	0	+ + -	0.29	0.96	0	- + -
Conversion of malonyl-CoA and acetyl-CoA to palmitate	0.523	0.614	0	+ + -	0.919	0.081	0	+ - -	0	1	0	- + -
Formation of Malonyl-CoA from Acetyl-CoA (liver)	0.415	0.586	0	+ + -	0.934	0.028	0	+ - -	0.583	0.583	0	+ + -
Generation of Cytoplasmic Acetyl CoA from Citrate	0.469	0.921	0	+ + -	0.934	0.039	0	+ - -	0.586	0.727	0	+ + -
3-oxooctadecanoyl-CoA (3-oxostearoyl-CoA) + NADPH + H+ => 3-hydroxyoctadecanoyl-CoA + NADP+	0.691	0.264	0	+ - -	0.784	0.31	0	+ - -	1	0	0	+ - -
arachidonate + CoASH + ATP => arachidonoyl-CoA + AMP + pyrophosphate + H2O [ACSL4]	0.265	0.974	0	- + -	0.475	0.467	0	+ - -	0.48	0.934	0	+ + -
Conversion of palmitic acid to palmitoyl-CoA	0.265	0.974	0	- + -	0.475	0.467	0	+ - -	0.48	0.934	0	+ + -
palmitate + CoASH + ATP => palmitoyl-CoA + AMP + pyrophosphate + H2O [ACSL3]	0.265	0.974	0	- + -	0.475	0.467	0	+ - -	0.48	0.934	0	+ + -
palmitate + CoASH + ATP => palmitoyl-CoA + AMP + pyrophosphate + H2O [ACSL5]	0.265	0.974	0	- + -	0.475	0.467	0	+ - -	0.48	0.934	0	+ + -
palmitate + CoASH + ATP => palmitoyl-CoA + AMP + pyrophosphate + H2O [ACSL6]	0.265	0.974	0	- + -	0.475	0.467	0	+ - -	0.48	0.934	0	+ + -

TABLE 5.11 – Valeurs de similarité et de particularité sémantiques mesurées sur BP, MF et CC pour chaque étape du métabolisme des acides gras, des triglycérides et des corps cétoniques entre *Homo sapiens* et *Gallus gallus*. Les motifs correspondants sont indiqué.

2.3 Interprétation

Les résultats obtenus pour les deux comparaisons présentées (Homme - souris et Homme - poule) montrent d'importantes différences. Sur le plan de la similarité, seuls deux des 289 réactions comparées entre l'Homme et la souris ont résulté en une similarité inférieure au seuil τ_{sim} pour la branche BP, et deux également pour la branche CC. Toutes les comparaisons faites sur MF ont eu un résultat supérieur à τ_{sim}. En revanche lorsque l'on compare l'Homme à la poule, près de la moitié des résultats de comparaison sur BP (105 sur 216) sont inférieurs à ce seuil. Ce nombre est important également sur CC (80) alors qu'il reste faible pour MF (13).

Si l'on avait utilisé uniquement une mesure de similarité sémantique, à l'image de la représentation proposée par la Figure 10 nous aurions fait face à deux limitations et l'analyse n'aurait pas pu être plus poussée. Premièrement, dans les cas de faible similarité, il est impossible de savoir à l'aide de la mesure similarité laquelle ou lesquelles des deux espèces a ou ont des spécificités. Deuxièmement, dans les cas de forte similarité, la mesure de similarité seule n'est pas capable de montrer si une des deux espèces a des fonctions particulières en plus des fonctions communes. C'est la mesure de particularité qui nous permet d'aller plus loin.

Les représentations qui convertissent en couleurs basées sur les composantes rouges, vertes et bleues (RVB) les valeurs de nos triplets de similarité et de particularités (Figures 12 et 18) permettent de se rendre compte de l'existence de particularités, aussi bien chez l'Homme que chez la souris et la poule. Cependant, ce type de représentation illustre une limite d'un modèle continu. En effet, dans la Figure 12, on compte environ autant de feuilles ayant un ton violet (soit un mélange de bleu et de rouge, signifiant une particularité de la souris parmi des gènes similaires) que de feuilles ayant un ton bleu-vert (soit un particularité de l'Homme parmi des gènes similaires). Or lorsque l'on discrétise les couleurs à l'aide de nos seuils de similarité τ_{sim} et de particularité τ_{par} pour obtenir la Figure 13, on ne compte plus que deux feuilles violettes indiquant une particularité de la souris parmi des gènes similaires contre 26 feuilles de couleur bleu-vert indiquant une particularité de l'Homme parmi des gènes similaires. Ces données correspondent aux valeurs que l'on trouve dans la table 5.2.

Concernant la poule, la plupart des particularités mesurées par rapport à l'Homme sont nulles. Trois sont supérieures à τ_{par} pour BP, deux pour MF et une pour CC. Il s'avère que ces six cas concernent trois réactions au total.

La souris a également de nombreuses valeurs de particularités nulles par rapport à l'Homme, mais moins que la poule. Elle n'a cependant pas plus de valeurs de particularité supérieures au seuil de particularité. En effet, on compte 3 valeurs de particularité pour la souris supérieures à τ_{par} pour BP, et trois pour CC, qui se trouvent être dans les trois mêmes réactions.

Le fait d'avoir un nombre réduit de cas semblables permet de les examiner de plus près ; c'était un but du système de seuils. Cela permet de regarder individuellement chacun des cas de valeur de particularité de la souris et de la poule supérieure à τ_{par}. Un fait intéressant concerne la réaction "*Arylsulfatase A hydrolyses sulfate from sulfatide to form cerebroside*", une réaction de la subdivision « métabolisme des glycosphingolipides » du métabolisme des sphingolipides. La souris comme la poule ont une valeur de particularité

supérieure à τ_{par} pour BP et CC par rapport à l'Homme. Dans les deux cas, il s'agit de motifs - + + pour BP et CC. Les Figures 23, 24 et 25 présentent respectivement la comparaison des annotations du gène ARSA entre l'Homme et la souris, l'Homme et la poule et la poule et la souris.

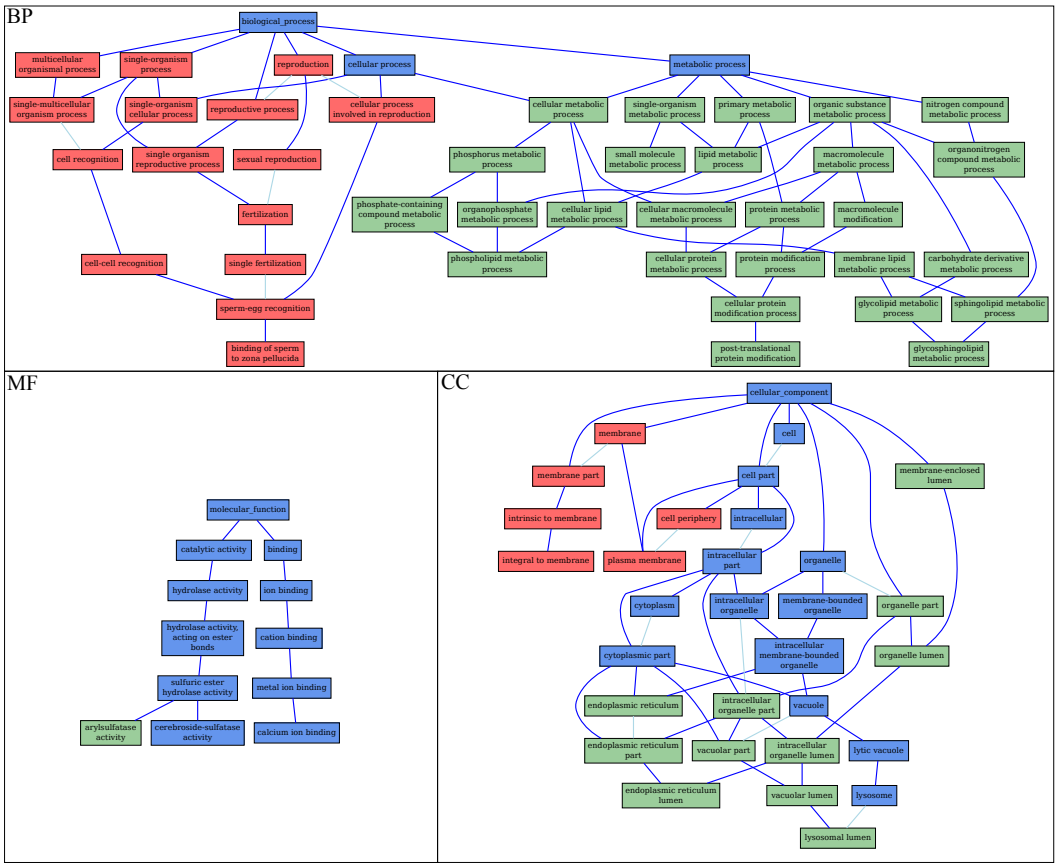


FIGURE 23 – Annotations GO du gène ARSA codant pour l'Arylsulfatase A chez l'Homme et la souris. Les termes GO en vert n'annotent que le gène humain, les termes en rouge n'annotent que l'orthologue murin et les termes en bleus sont communs.

L'annotation de la poule et celle de la souris sont très proches, d'où la différence observée dans les deux cas lors de la comparaison avec le gène humain. Malgré ces importantes différences d'annotations, les gènes ARSA de l'Homme, de la souris et de la poule sont tous dans le même groupe d'HomoloGene [1]. Les fonctions moléculaires de ce gènes sont bien conservées entre les espèces, cependant l'annotation laisse penser que les orthologues sont impliqués dans des processus biologiques différents et dans des composants cellulaires différents. Le terme "Binding of sperm to zona pellucida" est absent de l'annotation de l'Homme, cependant il a été démontré récemment que le gène ARSA chez l'Homme est impliqué dans ce processus [Xu et al., 2012]. Par conséquent l'annotation du gène humain devrait être corrigée, et l'on ne devrait plus observer de particularité pour la souris ou la poule par rapport à l'humain au moins au niveau de BP.

1. http://www.ncbi.nlm.nih.gov/homologene/20138

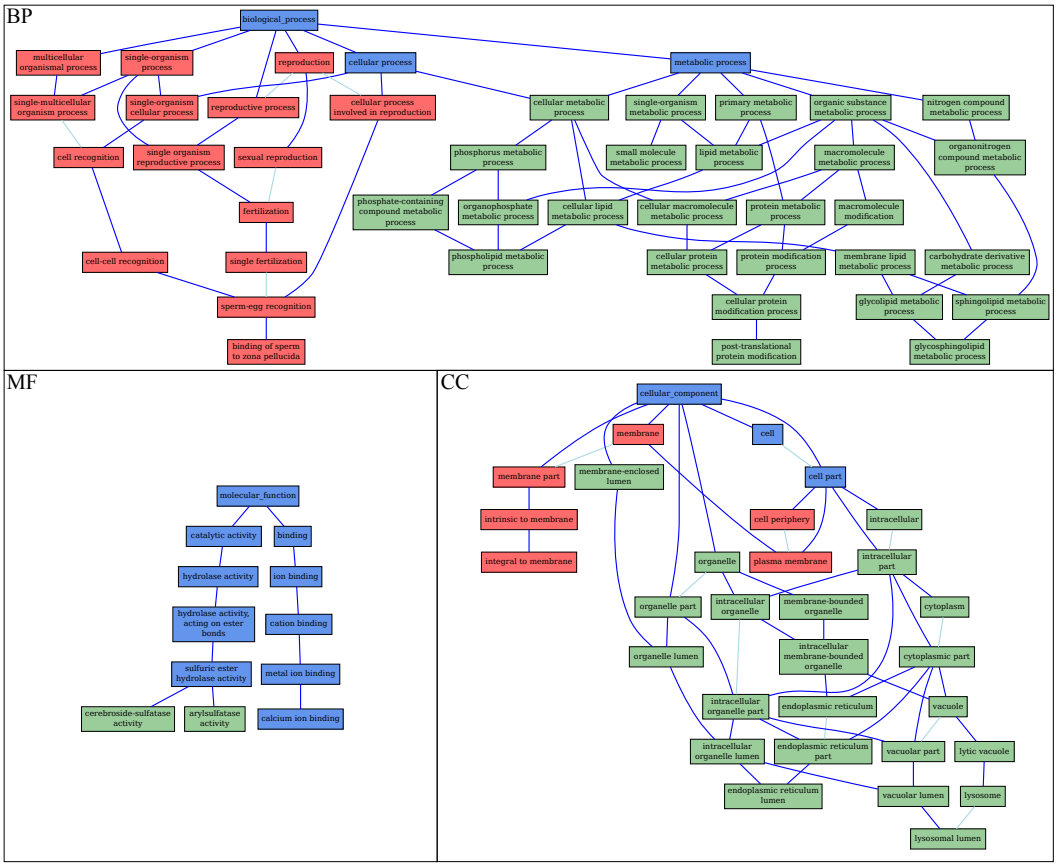


FIGURE 24 – Annotations GO du gène ARSA codant pour l'Arylsulfatase A chez l'Homme et la poule. Les termes GO en vert n'annotent que le gène humain, les termes en rouge n'annotent que l'orthologue de la poule et les termes en bleus sont communs.

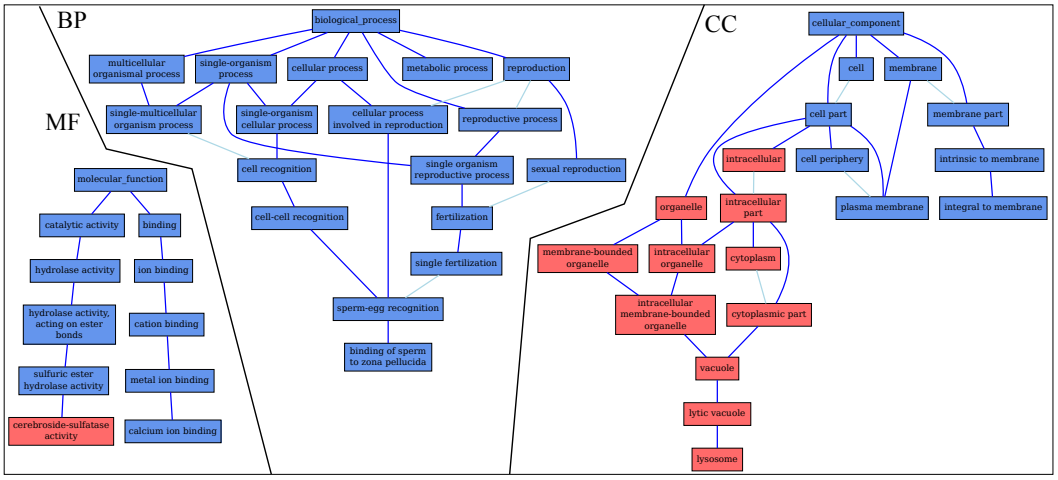


FIGURE 25 – Annotations GO du gène ARSA codant pour l'Arylsulfatase A chez la poule et la souris. Les termes en rouge n'annotent que l'orthologue murin et les termes en bleus sont communs.

Un deuxième cas parmi les particularités élevées mesurées chez la poule est celui de F1NC39 qui intervient dans la réaction "*PETA and CTP are condensed to CDP-ETA by PCY2*". D'après sa fiche UniProt[2], cette "Uncharacterized protein" est codée par le gène PCYT2, qui chez l'Homme code pour l'Ethanolamine-phosphate cytidylyltransferase qui catalyse justement cette réaction. Or selon HomoloGene, le gène PCYT2 de la poule n'est pas dans le même groupe que le PCYT2 de l'Homme. En effet, le gène PCYT2 de l'Homme est dans le goupe n°2143 avec des gènes codants pour des phosphate cytidylyltransferase d'autres espèces mais pas la poule[3] alors que le gène PCYT2 de la poule est dans le groupe n°56152 avec des gènes codants pour des Sirtuines, dont SIRT7 de l'Homme[4]. F1NC39 ayant une forte particularité à la fois sur BP, MF et CC et au vu de son annotation[5], on peut penser que ce gène n'est pas celui qui catalyse la réaction mentionnée chez la poule.

3 Biais et limites de la comparaison

Les approches développées et utilisées dans le cadre de cette thèse permettent plus de précision dans la comparaison inter-espèces de voies métaboliques, cependant il existe des biais et limitations dont nous allons discuter dans cette section.

3.1 Structure des voies métaboliques

La base de données de voies métaboliques parfaite n'existe pas. Une telle base disposerait de connaissances manuellement revues sur toutes les espèces susceptibles de nous intéresser et proposerait librement ses données dans un format standard (type BioPAX). Malheureusement comme dit dans le chapitre 2, chaque base de données a ses espèces de prédilection, parfois ses propres formalismes et n'est pas toujours libre d'accès. Pour s'assurer d'utiliser les meilleures données disponibles pour chaque comparaison de voies métaboliques, il faudrait donc intégrer les connaissances de plusieurs bases hétérogènes. Soh et al. [2010] ont fait état du faible niveau de cohérence, d'exhaustivité et de compatibilité entre ces bases de données. Cela rend donc cette tâche d'intégration difficile, longue et sans garantie de réussite.

Nous avons fait le choix de présenter ici des travaux basés sur Reactome, parce qu'il s'agissait de la seule base disponible ayant des données manuellement revues sur les trois espèces qui nous intéressaient. Cependant, le fait que Reactome soit avant tout basé sur les connaissances concernant l'Homme se ressent. Les données des autres espèces sont obtenues par un processus d'inférence validé manuellement. Il en résulte que sur toutes nos comparaisons structurelles des voies métaboliques, les seules différences constatées portaient sur la présence ou l'absence de quelques réactions. L'organisation générale des

2. http://www.uniprot.org/uniprot/F1NC39
3. http://www.ncbi.nlm.nih.gov/homologene/2143
4. http://www.ncbi.nlm.nih.gov/homologene/56152
5. http://www.bettembourg.fr/labo/go2graph/build.php?gene1=PCYT2&tax1=9031&gene2=PCYT2&tax2=9606

voies métaboliques est la même quelle que soit l'espèce considérée : celle établie pour l'Homme que l'on a acceptée pour les autres espèces de la base.

Nous avons pourtant le sentiment qu'il devrait y avoir des variations. Nous avons décrit les particularités du métabolisme des lipides chez les oiseaux en général et la poule en particulier dans le premier chapitre. Des éléments tels que le métabolisme quasi exclusivement hépatique de cette espèce, un système de transport particulier (absence de système lymphatique, absence d'apolipoprotéine E), un mécanisme de satiété sans leptine, devraient avoir un effet sur la structuration des voies métaboliques.

3.2 Annotations

La base de la comparaison sémantique de produits de gènes réside dans leurs annotations Gene Ontology. Comme les bases de données de voies métaboliques, cet aspect a ses limites.

3.2.1 Evidence codes

Nous avons vu que l'annotation d'un gène par un terme GO est associée à un "*Evidence code*". Le nombre important permet précisémment de savoir comment et pourquoi on a annoté tel gène avec tel terme GO. Cependant, encore une fois, acquérir des connaissances manuellement, prouver qu'elles sont vraies et pouvoir le retranscrire à travers une annotation de gène est un processus long et complexe. Il n'est possible que pour les espèces et les aspects de la biologie les plus étudiés. Les technologies à haut débit permettent de générer des péta-octets de données qui demanderaient des siècles de traitement afin d'être analysées manuellement. Par chance, des systèmes d'inférence automatique de plus en plus performants sont mis en place.

La limite de ce système est qu'afin d'éviter au maximum de faire une erreur en annotant automatiquement un gène, il faut se contenter d'utiliser des termes suffisamment généraux. Cela peut résulter en une annotation relativement vague commune à plusieurs orthologues. Dès lors, leur comparaison est biaisée. Même si l'on s'attend à trouver des particularités, parce qu'il est prouvé qu'il en existe (grâce à la littérature ou à des données propres à son laboratoire), si l'annotation d'un gène est à cent pour cent automatique, il y a bien peu de chance de les mettre en évidence.

Certes, ces annotations répondant au code IEA ne sont pas fausses comme on l'entend parfois, mais elles sont la plupart du temps bien peu informatives. A la lumière de la figure 8, il n'est pas surprenant d'avoir obtenu presqu'exclusivement des valeurs de particularité nulles pour la poule lors de nos comparaisons. Plus de 95 % d'annotations inférées automatiquement pour les gènes d'une espèce peuvent conduire à un grand nombre d'annotations mais à peu d'information au final.

La solution n'est sans doute pas dans l'espoir de voir un jour toutes les données revues manuellement, objectifs sans doute inatteignable, mais en l'amélioration des processus d'inférence afin d'obtenir par ce moyen une information plus riche.

3.2.2 Exhaustivité des annotations

Même lorsque l'on dispose de deux gènes correctement annotés pour lesquels on obtient des résultats de similarité et particularité intéressants (par exemple l'existance d'une forte particularité pour un gène malgré une forte similarité entre les deux), il n'est pas possible de valider formellement un gain ou une perte de fonction. En effet, le système d'annotation fonctionne selon un modèle de monde ouvert : ce n'est pas parce qu'un terme n'annote pas un produit de gène que ce dernier n'a pas la fonction décrite par ce terme. C'est une limite à toujours garder à l'esprit lorsqu'on interprète des résultats d'une mesure sémantique.

3.3 Comparaison de gènes par paires

Les méthodes que nous avons utilisées et développées permettent de comparer deux ensembles d'annotations, qui peuvent être celles de deux produits de gènes ou l'agglomération des annotations de plusieurs produits de gènes. Nous avons procédé ainsi dans notre comparaison des gènes intervenant dans les réactions du métabolisme des lipides lorsque qu'une de ces réaction pouvait faire intervenir plusieurs isozymes. Le biais provoqué par cette approche est que l'on n'a pas tenu compte de la possible redondance des annotations. En effet, certains termes GO plus fréquents que d'autres n'ont pas eu d'effet plus fort sur le résultat de nos mesures. A l'inverse, il est possible qu'un terme GO rare n'annotant qu'un seul gène ait provoqué une hausse de la particularité de tout l'ensemble alors que ce terme aurait pu être considéré comme anecdotique.

Il existe des méthodes de mesure de similarité capable de comparer plusieurs groupes d'annotations en même temps, cependant nous avons préféré nous focaliser sur une méthode qui ne travaille que sur deux gènes à la fois mais capable de fonctionner dans le cadre d'une comparaison inter-espèces.

4 Conclusion

Au travers de l'exemple de la comparaison du métabolisme des lipides chez l'Homme, la souris et la poule, nous avons vu l'intérêt d'aborder cette problématique sous les trois angles que sont la comparaison de la structure de la voie métabolique, la mesure de la similarité sémantique des produits de gènes présents à chaque étape, et la mesure de leur particularité. Ces trois approches sont complémentaires et leurs résultats peuvent être rassemblés dans un graphe présentant à la fois la structure de la voie métabolique, les points communs et les différences au niveau fonctionnel. L'utilisation d'un seuil de similarité et de particularité a permis de distinguer des cas potentiellement intéressants, qu'ils reflètent une possible réalité biologique (seulement "possible" car sous l'hypothèse d'un monde ouvert) ou une vraisemblable erreur dans une base de données.

Troisième partie

Autres applications

CHAPITRE 6

APPLICATION DES MÉTHODES SÉMANTIQUES À D'AUTRES PROBLÉMATIQUES

Dans ce chapitre, nous présentons les résultats obtenus en appliquant les compétences et méthodes développées au cours de cette thèse à d'autres problématiques, nécessitant elles-aussi des mesures sémantiques et/ou l'usage de gene ontology.

Nous avons ainsi développé un moteur de recherche de bibliographique appelé GO2PUB qui utilise les termes GO comme mots-clés pour interroger PubMed en procédant à une extention de requête avec les gènes annotés par le(s) terme(s) choisi(s). Ce travail est une extension d'un script développé en marge de mon stage de master 2 afin de trouver plus efficacement des références bibliographiques pertinentes pour la validation biologique des pistes que nous avions identifiées.

Nous avons également appliqué une mesure de similarité sémantique aux gènes de chaque groupe de la "*Duplicated Gene Database*" (DGD) développée par l'équipe Génétique et Génomique de l'UMR PEGASE. Le but de cette base de données est de classer les gènes dupliqués co-localisés par groupes et de fournir pour chaque groupe les valeurs de co-expression et de similarité sémantique intra-groupes. Ce travail a été réalisé pour neuf espèces.

Enfin, comme Gene Ontology constitue la base de notre matériel d'étude, nous nous sommes interrogés sur l'évolution de sa complexité. L'idée initiale était d'étudier si la somme des valeurs sémantiques (au sens de la mesure de Wang) des termes de GO fournissait une mesure intéressante de la complexité de l'ontologie. Les résultats ont montré qu'elle n'apportait pas grand chose aux mesures classiques de complexité de graphe (taille, connectivité et hiérarchie). Nous avons donc restreint l'article à l'étude de ces dernières mesures.

Sommaire

1 Développement d'une méthode et d'un outil de recherche bibliographique utilisant GO : GO2PUB

GO2PUB : Querying PubMed with semantic expansion of Gene Ontology terms

- Résumé -

Contexte : Le développement de méthodes d'analyses génétiques à haut débit crée un besoin d'outils capables d'interroger PubMed à la recherche d'articles pertinents relatifs au domaine d'étude. Avec la croisssance de PubMed, la recherche de bibliographie devient plus complexe et chronophage. Des outils de recherche ayant une bonne précision et un bon rappel sont nécessaires. Nous avons développé GO2PUB pour enrichir automatiquement des requêtes envoyées à PubMed avec des noms, symboles et synonymes de gènes annotés par un terme de Gene Ontology (GO) d'intérêt ou un de ses descendants dans le graphe de GO.

Résultats : GO2PUB enrichit les requêtes envoyées à PubMed sur la base de termes GO et de mots-clés choisis par l'utilisateur. Il compile les résultats et affiche le PMID, titre, auteurs, résumés et références bibliographiques des articles obtenus. Les noms, symboles et synonymes de gènes qui ont été utilisés en tant que mots-clés additionnels sont surlignés dans les résultats. GO2PUB est basé sur une expansion sémantique des requêtes envoyées à PubMed utilisant les relations d'héritage entre les termes du graphe de GO. La pertinence des résultats obtenus avec GO2PUB, GoPubMed et l'utilisation simple de PubMed a été évaluée par deux experts concernant trois requêtes sur le métabolisme des lipides. L'accord entre les experts était élevé (kappa = 0.88). GO2PUB a obtenu 69 % des articles pertinents, GoPubMed : 40 % et PubMed : 29 %. GO2PUB et GoPubMed ont obtenu 17 % de résultats communs, correspondant à 24 % du nombre total de résultats pertinents. 70 % des articles obtenus avec plus d'un outil se sont avérés être pertinents. En ce qui concerne les articles pertinents obtenus par un seul outil, 36 % des articles pertinents ont été trouvés seulement par GO2PUB, 17 % seulement par GoPubMed and 14 % seulement par PubMed.

Généralisation : Pour déterminer si ces résultats pouvaient être généralisés, nous avons généré 20 requêtes basées sur des termes GO pris au hasard ayant une granularité similaire à ceux des trois premières requêtes et nous avons comparé les proportions de résultats obtenus par GO2PUB et GoPubMed. Celles-ci étaient de respectivement de 77 % et 40 % pour les premières requêtes, et de 70 % et 38 % pour les requêtes basées sur des termes GO pris au hasard. Les deux experts ont aussi procédé à la vérification de pertinence des résultats de sept de ces vingt nouvelles requêtes (trois qui étaient relatives au métabolisme des lipides et quatre relatives à un autre domaine. L'accord entre expert était haut (0.93 et 0.8). Les performances de GO2PUB et GoPubMed étaient similaires à celles obtenues avec les trois premières requêtes.

Conclusion : Nous avons démontré que l'utilisation de gènes annotés soit par des termes GO d'intérêt, soit par un de leurs descendants permettait d'obtenir des articles pertinents non trouvés par d'autres outils. La comparaison de GO2PUB, basé sur une expansion sémantique, et GoPubMed, basé sur des techniques de "*text-mining*" a montré que ces outils sont complémentaires. L'analyse des requêtes générées aléatoirement suggère que les résultats obtenus pour le métabolisme des lipides peut être généralisé à d'autres processus biologiques. GO2PUB est disponible à `http://go2pub.genouest.org`.

JOURNAL OF BIOMEDICAL SEMANTICS

RESEARCH Open Access

GO2PUB: Querying PubMed with semantic expansion of gene ontology terms

Charles Bettembourg[1,2*], Christian Diot[2], Anita Burgun[1] and Olivier Dameron[1]

Abstract

Background: With the development of high throughput methods of gene analyses, there is a growing need for mining tools to retrieve relevant articles in PubMed. As PubMed grows, literature searches become more complex and time-consuming. Automated search tools with good precision and recall are necessary. We developed GO2PUB to automatically enrich PubMed queries with gene names, symbols and synonyms annotated by a GO term of interest or one of its descendants.

Results: GO2PUB enriches PubMed queries based on selected GO terms and keywords. It processes the result and displays the PMID, title, authors, abstract and bibliographic references of the articles. Gene names, symbols and synonyms that have been generated as extra keywords from the GO terms are also highlighted. GO2PUB is based on a semantic expansion of PubMed queries using the semantic inheritance between terms through the GO graph. Two experts manually assessed the relevance of GO2PUB, GoPubMed and PubMed on three queries about lipid metabolism. Experts' agreement was high (kappa=0.88). GO2PUB returned 69% of the relevant articles, GoPubMed: 40% and PubMed: 29%. GO2PUB and GoPubMed have 17% of their results in common, corresponding to 24% of the total number of relevant results. 70% of the articles returned by more than one tool were relevant. 36% of the relevant articles were returned only by GO2PUB, 17% only by GoPubMed and 14% only by PubMed. For determining whether these results can be generalized, we generated twenty queries based on random GO terms with a granularity similar to those of the first three queries and compared the proportions of GO2PUB and GoPubMed results. These were respectively of 77% and 40% for the first queries, and of 70% and 38% for the random queries. The two experts also assessed the relevance of seven of the twenty queries (the three related to lipid metabolism and four related to other domains). Expert agreement was high (0.93 and 0.8). GO2PUB and GoPubMed performances were similar to those of the first queries.

Conclusions: We demonstrated that the use of genes annotated by either GO terms of interest or a descendant of these GO terms yields some relevant articles ignored by other tools. The comparison of GO2PUB, based on semantic expansion, with GoPubMed, based on text mining techniques, showed that both tools are complementary. The analysis of the randomly-generated queries suggests that the results obtained about lipid metabolism can be generalized to other biological processes. GO2PUB is available at http://go2pub.genouest.org.

Keywords: Gene ontology, Semantic expansion, Query enrichment, PubMed

*Correspondence: charles.bettembourg@univ-rennes1.fr
[1] UMR936, INSERM, Université de Rennes 1, 2 av. Léon Bernard, F-35043 Rennes, France
[2] UMR1348, INRA, Agrocampus Ouest, 65 rue de Saint-Brieuc, F-35042 Rennes, France

Background

The development of high-throughput methods of gene analysis requires to deal with lists of thousands of genes while researchers were used to search the literature only for a few genes at a time. The information retrieval process becomes an increasingly difficult task and needs to be redesigned to provide literature concerning biological problems raised by the gene analyses.

PubMed is the most comprehensive public database of biomedical literature. It comprises more than 21 million entries for biomedical literature from MEDLINE, life science journals, and online books[a]. The typical PubMed user has to read several dozens to hundreds of abstracts to select the relevant ones. More than 4 million articles were added in the last 5 years [b].

A well defined query is important to retrieve as many relevant articles as possible with as few irrelevant ones as possible. Such a query is often more complex than the few loosely-coupled keywords used by most users. There is a need for automatic tools helping the users to build such complex queries that minimize silence and noise [1,2].

Although PubMed supports MeSH-based query expansion [3], other literature search tools have been developed [4-7] and evaluated [8]. These can be classified into three major approaches. The first approach, exemplified by tools like SLIM [9], is based on an intuitive interface to set some filters on PubMed queries in order to obtain a better precision than with the basic PubMed querying system. A good proficiency with PubMed *advanced search* brings similar results.

The second approach developed in SEGOPubMed uses a Latent Semantic Analysis (LSA) framework. It is based on a semantic similarity measure between the user query and PubMed abstracts [10]. The authors of SEGOPubMed state that the LSA approach outperforms the other approaches when using well-referenced keywords. Unfortunately, no implementation of SEGOPubMed is currently available. Moreover, this method requires that a corpus of well-referenced keywords be constituted and maintained before the search. Such a corpus is not available (in the biomedical domain) either.

The third approach is based on query enrichment using controlled vocabularies and ontologies. An ontology is a knowledge representation in which concepts are described both by their meaning and their relations to each other [11]. Ontologies are useful to find information relevant to a given topic, particularly through a query expansion process[12]. The automatic handling of the query complexity facilitates query formulation. Expanded queries applied to the web information retrieval show a systematic improvement over the unexpanded ones [13]. QuExT performs a concept-oriented query expansion to retrieve articles associated with a given list of genes symbols from PubMed and to prioritize them [14].

However, a frequent goal of gene-related analyses (e.g. transcriptomics) is to identify the genes with different expression across samples analyzed. Thereafter, scientists link their list of genes to more synthetic keywords and functions using Gene Ontology (GO) terms [15] associated to genes thanks to the Gene Ontology Annotation database [16]. At this stage of the gene-related analyses, the keywords to search the literature are not gene names anymore but GO terms. Therefore, tools querying litterature with GO terms seem appropriate. GoPubMed [17] uses a text extraction algorithm to mine PubMed abstracts with GO terms. It relies on a local string alignment to compare the GO terms and the abstracts. GoPubMed selects the abstracts containing at least a significant part of the semantic of the GO terms. However, GoPubMed does not follow GO strict rules conveying the semantics of terms. If the annotation of a gene product gp by a Gene Ontology term t is true, then the annotation of gp by any parent of t is equally true [16]. All transitive relation (is a, part of) have to be followed to retrieve these parents. As GoPubMed does not follow this rule, its recall decreases whenever inferences about gene annotations yield new relevant results [18]. None of the existing tools supports a combination of semantics-based and of synonym-based PubMed query enrichment.

In this study, we hypothesized that the name of the genes annotated by a GO term of interest or one of its descendants can be used as keyword in gene-oriented PubMed queries. The descendants of a GO term are defined according to the Gene Ontology specifications of reasoning about relations[c]. The genes annotated with GO terms are provided by the Gene Ontology Annotation database.

In our system GO2PUB, we propose a new approach that considers not only the genes annotated with a GO term of interest, but also those annotated by a descendant of this GO term, complying with the semantic inheritance properties of GO. GO2PUB's user inputs a list of GO terms of interest, one or more species, and a list of keywords. It generates a PubMed query with the names, symbols and synonyms or aliases of these genes, the species and the keywords and processes PubMed results.

We performed a qualitative relevance study on our domain of expertise using three queries related to lipid metabolism. Because GO2PUB and GoPubMed both use GO terms as input we wanted to confront the results from these tools. For each query, we compared GO2PUB results with those of the original GoPubMed and of GoPubMed after having manually-generated the semantic expansion of the GO terms. In addition, we submitted similar queries to PubMed as it is the reference literature search tool. Two experts manually determined the relevance of all the articles. We computed the precision, relative recall and F-score of GO2PUB, GoPubMed and PubMed. In order

to determine if the results of the qualitative study could be generalized, we then performed a study on twenty randomly-generated queries. This study focused on the number of common results and tool-specific results. We also analyzed the relevance of seven of these twenty random queries.

Results

Qualitative study

In order to evaluate GO2PUB's relevance and to compare it with GoPubMed, we assessed three queries (Q1, Q2 and Q3) about biological processes related to lipid metabolism and including different GO terms, species and MeSH terms. We submitted our queries to GoPubMed using the same keywords and tags. As GoPubMed only considers the GO term(s) provided by the user and ignores the inheritance rules of Gene Ontology, we also expanded queries manually then submitted them to GoPubMed. Our GoPubMed queries were composed of the GO term(s) and all its descendants separated by "OR", plus MeSH keywords. This ensured the closest comparison possible. We also constructed the PubMed queries as close as possible to our GO2PUB queries.

Relevance criteria

The role of GO2PUB is to retrieve literature about gene functions summarized by a GO term thanks to a gene analysis process. We analyzed the results of GO2PUB, GoPubMed and PubMed queries according to the following criteria. We considered that a relevant article had to describe at least one gene product occurring in the chosen domain of interest for the selected species. The gene product's description has to focus on its role, its interactions, and how and when it is activated.

For each query Q1, Q2 and Q3, the results obtained by the different tools were mixed for a blind selection by a biologist, CD and by a bioinformatician, CB. This ensured that the reviewers did not know which tool(s) retrieved the articles. The final list of relevant articles is the union of the two reviewers' lists.

Relevance measurement

For each query and tool, we computed the precision, recall and F-score. Computing the recall for each query is impossible because it would require to know all the relevant articles available in Medline. As it is possible that some of these articles were missed by all three tools, recall was defined as relative to all relevant articles obtained by at least one of the tools.

Figure 1 presents the reviewers' selections of relevant articles among all the results of the qualitative study. Most of the relevant articles were found in the intersection of the two selections. Reviewers agreed on 35 relevant and 113 irrelevant articles while selecting separately 3 and 4 articles as relevant. Additional files 1, 2 and 3 provide the experts' selections for Q1, Q2 and Q3.

We used Cohen's kappa coefficient as a statistical measure of inter-rater agreement [19]. The value obtained was 0.88, which corresponds to an almost perfect agreement [20].

Query Q1: Lipogenesis in chicken liver

For our first query in GO2PUB, we used "Lipid biosynthetic process" (`GO:0008610`) as GO term, "Gallus gallus" as species, "Liver" as Major Topic and "Lipid Metabolism" as MeSH keyword, and we considered the articles published in the last five years. We ran query Q1 on GO2PUB using the [BASICq], [MeSHq] and [ORq] options described in method section. "Lipid biosynthetic process" has 243 descendants in the GO graph. The mean number of edges to reach the root of the ontology from this term was 3.5.

Additional file 4 contents the results obtained by GO2PUB for query Q1. Results are formatted for a quick access to information. Each citation obtained from PubMed is listed; the title, authors, date, abstract, journal, PMID and MeSH terms are displayed. The name, symbol and synonyms of gene annotated by the GO term(s) are highlighted in the title and abstract.

The query Q1 formulated for GoPubMed included *"lipid biosynthetic process"[go] AND Chickens[mesh] AND Liver[majr] AND "Lipid Metabolism"[mesh] AND last5years[time]*. This is the "standard" query for GoPubMed. We also formed the manually-expanded version of this query by adding the descendants of "lipid biosynthetic process" separated by "OR". It should be noted that 47 of the 243 terms generated by the semantic expansion of "Lipid biosynthetic process" generated a GoPubMed error and had to be ignored. For example, one of the descendants of "Lipid biosynthetic process" is "Regulation of phospholipid biosynthetic process" (`GO:0071071`), which is a relevant descendant term. When querying GoPubMed with this GO term, we obtained an error: "Your query could not be understood: Can't find a term regulation of phospholipid biosynthetic process".

The PubMed equivalent query for Q1 was "Chickens liver lipogenesis", which PubMed interpreted as *("chickens"[mesh] OR "chickens"[all]) AND ("liver"[mesh] OR "liver"[all]) AND ("lipogenesis"[mesh] OR "lipogenesis"[all])*.

Figure 2 presents Venn diagrams comparing the results obtained with PubMed, GoPubMed (after manual expansion) and GO2PUB for Q1. Figure 2A presents the raw results. Although queries as similar as possible were issued to the three tools, the resulting sets of articles had little overlap. Figure 2B presents the repartition of the relevant articles. Most of the relevant articles were identified by GO2PUB. Of note, most of the articles retrieved by at

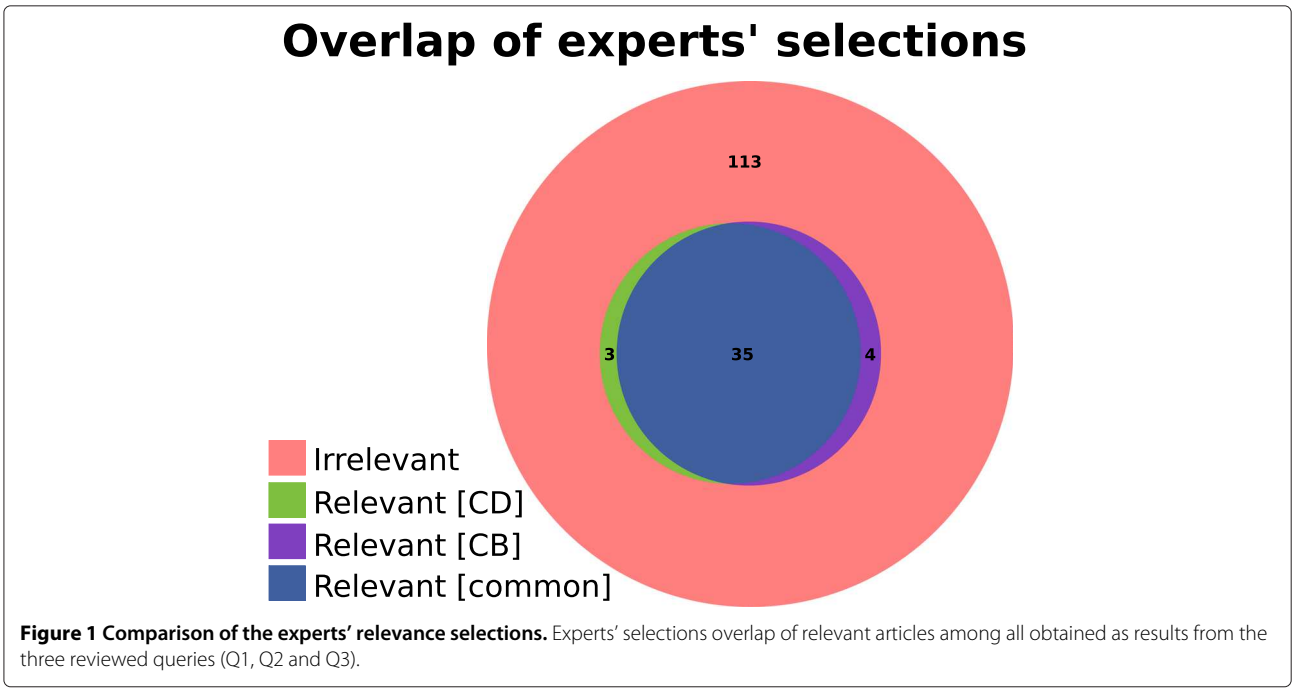


Figure 1 Comparison of the experts' relevance selections. Experts' selections overlap of relevant articles among all obtained as results from the three reviewed queries (Q1, Q2 and Q3).

least two tools (overlaps in Figure 2A) were found to be relevant (overlaps in Figure 2B).

Table 1 presents the precision, relative recall and F-score for each tool. GO2PUB had a better precision and relative recall than GoPubMed and Pubmed. Regarding GoPubMed, there was no difference between the "standard" and "expanded" results.

Query Q2: Lipid transport in human blood

In our second reviewed query in GO2PUB, we used "Lipid transport" (GO:0006869) as GO term, "Homo sapiens" as species, "Blood" as Major Topic and "Lipid Metabolism" as MeSH keyword, and we considered the articles published in the last five years. "Lipid transport" has 109 descendants in the GO graph. The mean number of edges to reach the root of the ontology from this term was 4.3.

We ran equivalent queries on GoPubMed ("standard" and "expanded" versions) and PubMed. 46 of the 109 terms generated by the semantic expansion of "Lipid Transport" generated a GoPubMed error and had to be ignored.

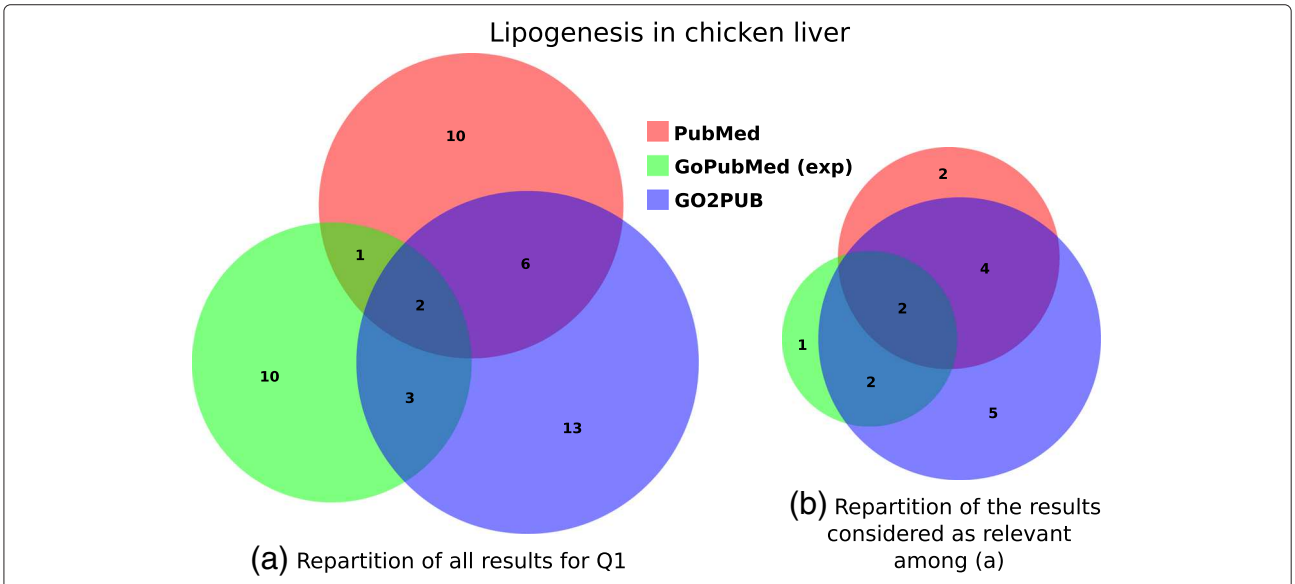


Figure 2 Comparison of the PubMed, GoPubMed and GO2PUB results for query Q1. (**a**) displays the repartition and intersections of these results. (**b**) displays the repartition and intersections of the results considered as relevant.

Table 1 Measures for query Q1

	PubMed	GoPM (std)	GoPM (exp)	GO2PUB
(a) Number of results	19	16	16	24
(b) Relevant among (a)	8	5	5	13
Precision	0.421	0.313	0.313	0.542
Relative Recall	0.5	0.313	0.313	0.813
F-score	0.457	0.313	0.313	0.650

Values of precision, relative recall and F-score using PubMed, GoPubMed without (GoPM std) or with (GoPM exp) manual expansion and GO2PUB search tools for query Q1 about lipogenesis in chicken liver. Values are calculated from (a) and (b) lines using a total number of relevant results of 16.

As there is no MeSH term for "lipid transport", we searched it in titles and abstracts on PubMed. The PubMed query was: *"lipid transport"[TIAB] AND (("blood"[Subheading] OR "blood"[All Fields] OR "blood" [MeSH Terms]) AND ("humans"[MeSH Terms] OR "humans"[All Fields] OR "human"[All Fields])) AND ("2006/03/28"[PDat] : "2011/03/28"[PDat])*.

Figure 3 presents the results obtained by PubMed, GoPubMed (after manual expansion) and GO2PUB for Q2. As observed for query Q1, the majority of the results were tool-specific. PubMed yielded 45 articles, none of which were retrieved by GO2PUB nor GoPubMed while there was an overlap between GO2PUB and GoPubMed results. Considering only GoPubMed and GO2PUB, most of the results were specific to one tool or the other and few were obtained by both tools (Figure 3A). Three of the four common articles between GoPubMed and GO2PUB were relevant (Figure 3B). GO2PUB yielded half of GoPubMed relevant results while having an important specific relevant results set. Only 2 article on 45 yielded by PubMed were relevant.

Table 2 presents precision, relative recall and F-score for each tool. GO2PUB has a slightly lower precision than GoPubMed (standard and after manual expansion) but better relative recall and F-score. For GoPubMed, there was no difference between "standard" and "expanded" results.

Query Q3: Regulation of lipase activity in human cell membrane

Our third query in GO2PUB used "Regulation of lipase activity" (`GO:0060191`) as GO term, "Homo sapiens" as species and "Cell Membrane" and "Lipid Metabolism" as MeSH keywords, and considered the articles published in the last ten years. "Regulation of lipase activity" has 35 descendants in the GO graph. The mean number of edges to reach the root of the ontology from this term was 5.25.

We ran equivalent queries on GoPubMed ("standard" and "expanded" versions) and PubMed. 16 of the 35 terms generated by the semantic expansion of "Regulation of lipase activity" generated a GoPubMed error and had to be ignored.

The PubMed query was composed of the keywords "regulation", ("lipase" AND "activity"), "human" and ("cell" AND "membrane").

Figure 4 presents the results obtained by PubMed, GoPubMed (after manual expansion) and GO2PUB. Figure 4A shows a larger set of results for GO2PUB compared to

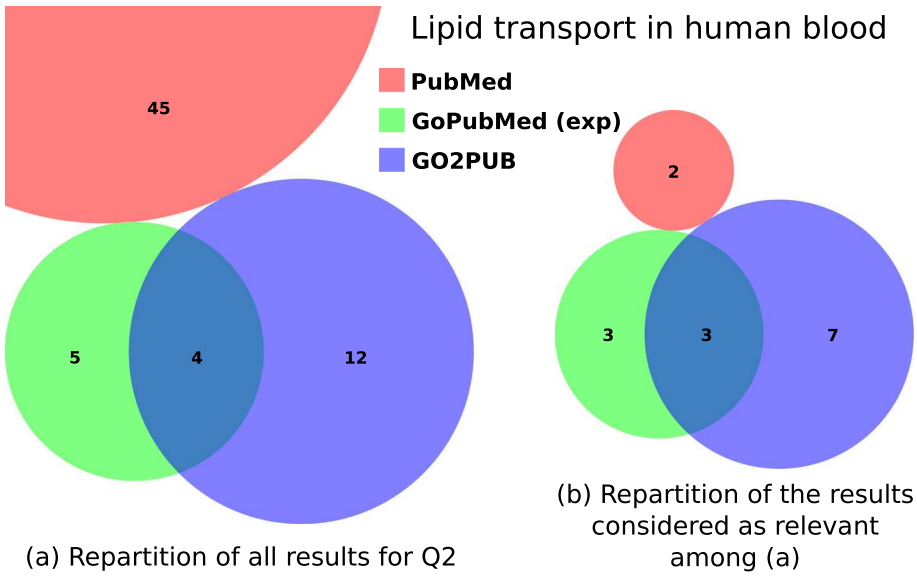


Figure 3 Comparison of the PubMed, GoPubMed and GO2PUB results for query Q2. (**a**) displays the repartition and intersections of these results. (**b**) displays the repartition and intersections of the results considered as relevant.

Table 2 Measures for query Q2

	PubMed	GoPM (std)	GoPM (exp)	GO2PUB
(a) Number of results	45	9	9	16
(b) Relevant among (a)	2	6	6	10
Precision	0.044	0.667	0.667	0.625
Relative Recall	0.133	0.4	0.4	0.667
F-score	0.067	0.5	0.5	0.645

Values of precision, relative recall and F-score using GoPubMed without (GoPM std) or with (GoPM exp) manual expansion and GO2PUB search tools for query Q2 about lipid transport in human blood. Values are calculated from (a) and (b) lines using a total number of relevant results of 15.

Table 3 Measures for query Q3

	PubMed	GoPM (std)	GoPM (exp)	GO2PUB
(a) Number of results	23	6	8	24
(b) Relevant among (a)	2	5	6	6
Precision	0.087	0.833	0.75	0.25
Relative Recall	0.182	0.455	0.545	0.545
F-score	0.118	0.588	0.632	0.343

Values of precision, relative recall and F-score using PubMed, GoPubMed without (GoPM std) or with (GoPM exp) manual expansion and GO2PUB search tools for query Q3 about regulation of lipase activity in human cell membrane. Values are calculated from (a) and (b) lines using a total number of relevant results of 11.

GoPubMed (24 and 8, respectively), but we can see in Figure 4B that most of these results are irrelevant. As observed for query Q2, none of the PubMed results were retrieved by GO2PUB nor GoPubMed while there was an overlap between GO2PUB and GoPubMed results. Only 2 articles on 23 identified by PubMed were relevant.

Table 3 presents precision, relative recall and F-score for each tool. GO2PUB has a relative recall equivalent to GoPubMed's and a lower precision and F-score. For GoPubMed, the "manually-expanded" results have a higher relative recall and F-score and a lower precision than the "standard" ones. We observed again a discrepancy between PubMed and the other tools, with a lower precision, a lower relative recall, and consequently a lower F-score for PubMed.

Generalization study

In order to determine whether previous results are representative of GO2PUB's performances, we performed a generalization study on twenty randomly-generated queries. We compared the profile of the results obtained by GO2PUB and GoPubMed in this generalization study with those obtained in the qualitative study. This profile depends on the average size of the sets of articles. The following proportions were calculated on the result set constituted by all GoPubMed and GO2PUB results. In the qualitative study, GO2PUB yielded 21.33 articles on average, which represented 77.1% of the total. GoPubMed yielded 11.0 articles on average, which represented 39.8% of the total. There were 4.67 articles on average in the set of common articles, which represented 16.9% of the total.

We built queries following the pattern: "a random GO term + a species (mouse) + a publication date limit (2011) + a keyword (the GO term name)". To be coherent with our qualitative study, we randomly selected twenty GO terms among all Biological Process terms having a granularity similar to those of the three GO terms used in the qualitative study. We assumed that the granularity of

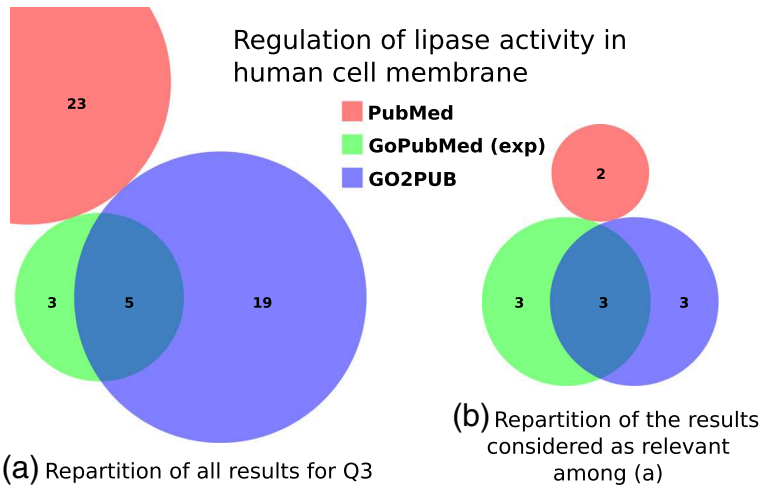


Figure 4 Comparison of the PubMed, GoPubMed and GO2PUB results for query Q3. (**a**) displays the repartition and intersections of these results. (**b**) displays the repartition and intersections of the results considered as relevant. The GoPubMed displayed set is the one that uses the manual semantic expansion.

a term depends on the mean length of its path to the root, and its number of descendants. Each GO term of the generalization study had a mean path length to the root between 3.5 and 5.25 edges and had between 35 and 244 descendants. As we could not add a MeSH keyword in relation with the random GO term of each query, we simply added the name of this GO term. This keyword was added in the free field for GO2PUB and without [go] tag for GoPubMed. We submitted these queries to GO2PUB and to GoPubMed.

Figure 5 presents the sets of articles obtained by GO2PUB and GoPubMed for these queries. GO2PUB yielded 46 articles on average (min 6, max 189) compared to 21.33 on the qualitative study. They represented 70.4% of the total number of articles (77.1% in the qualitative study). GoPubMed yielded 25.1 articles on average (min 2, max 88) compared to 11.0 on the qualitative study. They represented 38.4% of the total number of articles (39.8% in the qualitative study). There were 5.75 articles on average (min 0, max 59) in the common set. They represented 8.8% of the total (16.9% in the qualitative study). The profile of these results is close to the qualitative study one.

We studied the relevance of the results from seven queries picked out among the twenty queries of the generalization study. Out of the seven queries, three were chosen because they were in our reviewers' domain of expertise: "cellular lipid catabolic process" [GO:0044242], "isoprenoid biosynthetic process" [GO:0008299] and "phospholipid biosynthetic process" [GO:0008654]. Cohen's kappa was 0.9345. We picked randomly four additional queries about "RNA transport" [GO:0050658], "tetrapyrrole metabolic process" [GO:0033013], "xenobiotic metabolic process" [GO:0006805] and "organelle fusion" [GO:0048284]. Cohen's kappa remained high for these four queries (0.797) in spite of them being out of our reviewers' domain of expertise. Table 4 presents the number of results, precision, relative recall and F-score respectively for the three lipid-related queries and the other four queries of the generalization study. Results are similar to those observed in the qualitative study. The resulting sets of articles had little overlap. Moreover, each tool yielded relevant results ignored by the other, with important variation of performances among queries.

Discussion

Our goal was to develop a tool that uses the knowledge from the Gene Ontology (GO) and its annotations for generating semantically-expanded gene-related PubMed queries. Indeed, there is no [GO] tag for a search in PubMed.

The qualitative study showed that both GO2PUB and GoPubMed retrieved relevant articles ignored by PubMed. For the query Q1 about lipogenesis in chicken liver, 26 of the 35 articles (8 of 14 relevant) returned by either GO2PUB or GoPubMed were ignored by PubMed. Conversely, 9 of the 19 articles (6 out of 8 relevant) returned by PubMed were also returned by either GO2PUB or GoPubMed. For Q2 and Q3, the set of articles returned by PubMed was disjoint from both GO2PUB and GoPubMed results. PubMed identified only 4 relevant articles not yielded by GO2PUB nor GoPubMed for these 2 queries.

Overall, GO2PUB performed better than GoPubMed and PubMed. Both GoPubMed and GO2PUB and to a lesser extend PubMed yielded relevant articles ignored by

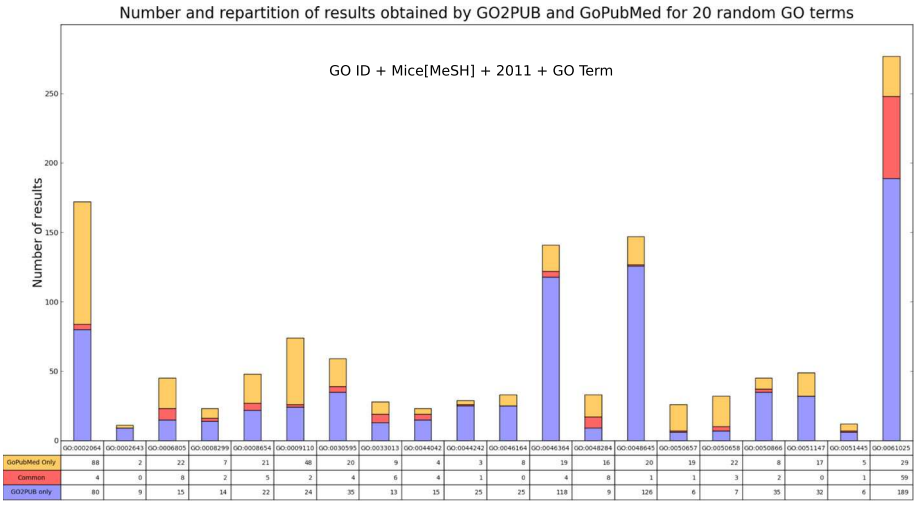


Figure 5 Repartition of results obtained by GO2PUB and GoPubMed for twenty general GO queries without free keyword. This diagram presents the sets of results obtained by GoPubMed and GO2PUB for twenty general queries. These queries were built with a random GO term having a granularity similar to that of GO terms of the qualitative study.

Table 4 Measures for seven generalization queries

(A) Lipids	**GO:0044242**		**GO:0008299**		**GO:0008654**	
Tool	GPM	G2P	GPM	G2P	GPM	G2P
(a) Number of results	4	26	9	16	25	27
(b) Relevant among (a)	3	20	1	2	5	11
(c) Total relevant	22		3		12	
(d) Common results	1		2		5	
(e) Relevant among (d)	1		0		4	
Precision	0.750	0.769	0.111	0.125	0.200	0.407
Relative Recall	0.136	0.864	0.333	0.667	0.417	0.917
F-score	0.231	0.814	0.167	0.211	0.270	0.564

(B) Other	**GO:0050658**		**GO:0033013**		**GO:0006805**		**GO:0048284**	
Tool	GPM	G2P	GPM	G2P	GPM	G2P	GPM	G2P
(a) Number of results	25	10	15	19	30	23	24	17
(b) Relevant among (a)	7	2	3	3	17	14	10	9
(c) Total relevant	9		6		26		16	
(d) Common results	3		6		7		8	
(e) Relevant among (d)	1		2		4		4	
Precision	0.280	0.200	0.200	0.158	0.567	0.609	0.417	0.529
Relative Recall	0.875	0.250	0.600	0.600	0.680	0.560	0.625	0.563
F-score	0.424	0.222	0.300	0.250	0.618	0.583	0.500	0.545

Results' sets sizes and values of precision, relative recall and F-score using GoPubMed and GO2PUB search tools for seven random queries: three lipid-related queries (part A) and four queries about other topics (part B). Values of precision, relative recall and F-score are calculated from (a), (b) and (c) lines. (d) and (e) lines provide GO2PUB and GoPubMed common results' sets sizes and the number of common relevant results.

the others. The discrepancy observed between PubMed and the other tools is probably due to the absence of a [GO] search field tag in PubMed. GO2PUB performance varied among the queries. For two queries (Q1, Q2) of the qualitative study, GO2PUB yielded most of the relevant articles and had therefore the highest relative recall value while its precision was slightly lower than that of GoPubMed. Consequently, GO2PUB had the best F-score. For Q3, GO2PUB yielded as many relevant articles as GoPubMed but had a higher noise proportion. GO2PUB had a slightly better relative recall than GoPubMed, but its precision was much lower. Consequently, GoPubMed had the best F-score. We can also notice that for Q3, the query expansion on GoPubMed improved its performances with a better relative recall and F-score at the cost of a small loss of precision. We observed similar results on the seven queries of the generalization study for which we assessed the relevance.

GO2PUB performs a semantic expansion of the GO terms of interest complying with the semantic inheritance through the GO graph before retrieving the corresponding genes to enrich the query. All the results of GO2PUB presented here were obtained using the concept of query expansion. During the development of GO2PUB, we also ran queries without this expansion. We obtained empty or very small sets of results.

Using the semantic inheritance properties of the GO graph is useful. The more descendants a GO term has, the more relevant results GO2PUB yields. GO2PUB performance decreased from Q1 to Q3. For Q1, "lipid biosynthetic process" has 243 descendants and annotates 646 genes for human and 145 genes for chicken. For Q2, "lipid transport" has 109 descendants and annotates 253 genes for human and 63 genes for chicken. For Q3, "regulation of lipase activity" has 35 descendants and annotates 168 genes for human and 18 genes for chicken. The more descendants a GO term has, the more genes it is likely to annotate. Moreover, Q1 concerned chicken, which is less annotated than human. On less annotated species, the annotations focus on the major genes. This explains why GO2PUB yields a high proportion of relevant articles.

Concerning Q3, GO2PUB had a low precision compared to GoPubMed. Genes annotated with GO terms on regulation usually have many additional functions. Consequently, the articles about genes annotated by "regulation of lipase activity" searched in Q3 may also describe the other functions of these genes. To obtain a better precision

in this case, we suggest to further specify the query with a MeSH term or a free keyword.

GoPubMed does not follow the semantic inheritance properties of GO. We manually expanded GoPubMed queries and compared it to GO2PUB. The added value of semantic expansion was null for Q1 and Q2, and important for Q3 (+33%). So query expansion is a built-in functionality in GO2PUB, and would be a valuable extension for GoPubMed. In GoPubMed results, a "missing term" error occurred for 19% of the expanded set of GO terms for Q1, 42% for Q2 and 44% for Q3. We assume that the benefits of query expansion on GoPubMed might be higher when considering the articles related to these currently omitted GO terms.

In order to verify whether the results from the qualitative study on lipid metabolism could be generalized to other domains, we submitted twenty randomly-generated queries to GO2PUB and GoPubMed. Each query contained a random GO term of a granularity similar to that of the terms used in the qualitative study. The proportion of articles returned by GO2PUB was 70.4%, the one of GoPubMed was 38.4% and the proportion of articles returned by both was 8.8%. These proportions were respectively 77.1%, 39.8% and 16.9% for the qualitative study. We assume that the difference of proportions between the qualitative and the generalization studies can be attributed to Q1, Q2 and Q3 being more specific because of the use of MeSH keywords. The seven queries of the generalization study presented relevances similar to those observed in the qualitative study.

GO2PUB seems less suited for queries involving either general GO terms or GO terms with few or no descendants. Indeed, with general GO terms, GO2PUB considers a lot of descendants, and therefore a lot of genes. We expect this to increase the noise as some of the genes will be irrelevant. Conversely, GO terms having few or no descendants are associated with few genes. We do not expect semantic expansion to benefit these highly specific queries yielding only a few PubMed results.

As most of the results obtained by GO2PUB and GoPubMed are relevant in the qualitative study and in the generalization study, the intersection of GoPubMed and GO2PUB results decreases noise. As each tool yields relevant articles ignored by the other, the union of their results also decreases silence.

Conclusion

GO2PUB brings relevant results ignored by GoPubMed (9 GO2PUB' specific results for Q1, 7 for Q2 and 3 for Q3) even when adding a manual query expansion for GoPubMed. Conversely GoPubMed text mining approach finds relevant articles ignored by GO2PUB (1 GoPubMed' specific result for Q1, 3 for Q2 and 3 for Q3). This demonstrates GO2PUB relevance and its complementarity with GoPubMed for our domain of interest. The generalization analysis shows that a similar profile of results is obtained using random queries, especially when using keywords for narrowing the queries. This suggests that the results of the qualitative study can be generalized.

Resources and methods

Resources

The files from GO[d] and GOA[e] used in our study were downloaded in March 2011. We used the "term" and "term2term" tables of GO for the automatic semantic expansion of GO2PUB and the manual expansion of GoPubMed queries. We used species specific GOA tables to retrieve for each species of interest the gene names annotated by the provided GO terms. These tables allowed us to build queries about seven different species[f]. Since June 2011, GO2PUB uses the Uniprot-GOA table instead of the species-specific tables, allowing researchers to mine the literature about more than 2000 different species. Additionally, this table is more complete than the species-specific tables used previously.

All the queries were submitted to GO2PUB, PubMed and GoPubMed on 28th March 2011. Synonyms and aliases of genes used in GO2PUB were provided by the current version of EntrezGene.

We represented the overlap of the different tools results using Venn diagrams generated by BioVenn [21].

Methods

GO2PUB query building

GO2PUB creates an expanded PubMed query with the name, symbol and synonyms of genes annotated by one or several GO terms provided by the users, for one or several species. Figure 6 presents the process. The users provide one or several GO terms and species. To further restrict their query, they can also provide as many MeSH terms keywords as wanted. Furthermore, a "free text" field supports the use of all the other PubMed tags, like [Author], [Journal], etc., and keywords from MeSH terms or free text.

The first part of each query involves one or more GO terms. The users can enter either the name or the identifier of the GO terms. These terms are suggested when the users start to fill the field. The exact GO term is suggested if the users provide one of its GO synonyms. For example, GO2PUB will search for "lipid biosynthetic process" if the users provide "lipogenesis". When two or more GO terms are entered, GO2PUB makes the union of them ("OR" connector).

Then, the users select one or several species using a name (common or scientific names and their synonyms are allowed) or a NCBI taxon code[g]. In this case, the users can choose to join them (using "OR") or intersect them (using "AND"). Logical connectors "AND" and

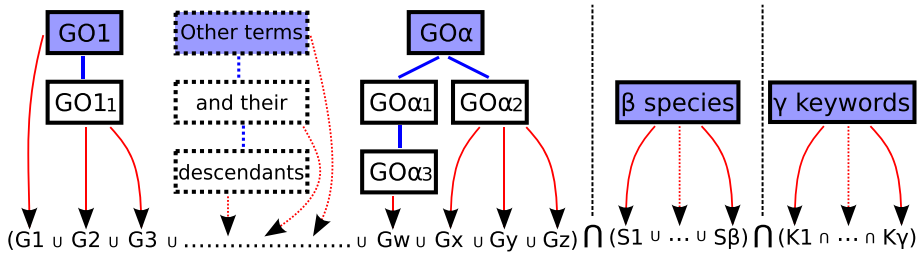


Figure 6 GO2PUB Query composition process using the parameters provided by the user. (1) the initial α GO terms (purple boxes) are enriched by their descendants. (2) the genes (here noted G1 to Gz) annotated by the GO terms are retrieved. (3) the query is composed using the names, symbols and synonyms of the genes, the β species (S) and the γ MeSH or free keywords (K).

"OR" are set by default to make the union of species and intersection of keywords, but this can be modified.

Next, the users can enter additional MeSH terms to specify their query. MeSH terms associated to the articles by PubMed are not all of same importance, some of them being classified as "Major topic" (MAJR). We can qualify each keyword as a simple MeSH term or a Major topic. Again, the users can specify the connector between keywords.

At this point, the users have built a simple GO2PUB query. We call this query [BASICq]. The system supports three modifications for [BASICq] for studying if minor changes bring additional relevant results.

The first modification ignores MAJR qualifiers and searches all keywords in PubMed [MeSH] tag. As MAJR terms are also MeSH terms, articles associated to them will still be found. We call this query [MeSHq].

The second modification replaces "AND" connectors between keywords by "OR" connectors. However, as it can return substantially more results with a lot of noise, all keywords in this additional query are tagged with MAJR. Species, normally searched in MeSH, are also tagged with MAJR. We call this query [ORq].

The third modification ignores MeSH and MAJR keywords, and tags species with MAJR. This option must be used carefully because it can yield several hundreds of results if the search topic is too large. It is of interest only for very narrow topics if the users do not obtain enough results with the other types of queries. We call this query [NOKq].

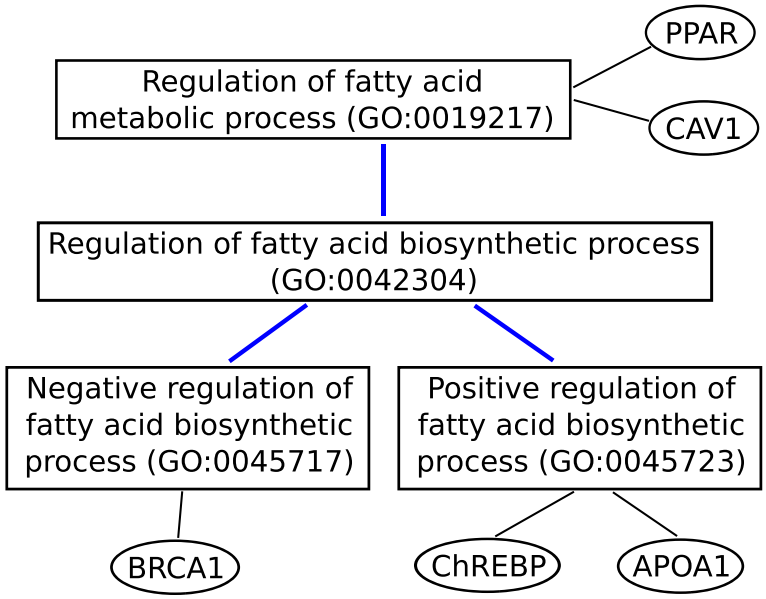


Figure 7 Keyword semantic enrichment. For a literature search about the regulation of fatty acid metabolic process, we want to enrich the query with the associated genes. The two genes PPAR and CAV1 are directly annotated by the GO term "Regulation of fatty acid metabolic process" (`GO:0019217`). However, Gene Ontology inheritance properties say that every term inherits the meaning of all its ancestors. Consequently, genes annotated by at least one descendant of the original term (BRCA1, ChREBP and APOA1) should also be considered.

Last, GO2PUB proposes three additional options.

The first option sets limits on the publication year.

The second option proposes an exhaustive search of the official synonyms of gene names. It searches Entrez gene[h] for all the known synonyms for a gene. Since authors sometimes use synonyms that are absent in the GOA database in their articles, this option allows the users to build more complete PubMed queries in order to obtain more relevant results.

The third option toggles the display of the MeSH table associated with each article.

Query rewriting using semantic expansion

Semantic expansion consists in following the semantic inheritance through the GO graph in order to also consider all the descendants of the GO terms specified by the users. Then, the process retrieves the gene names annotated with these terms.

GO2PUB uses these gene names and their synonyms as additional keywords for PubMed queries. Figure 7 shows that the expansion identifies five genes associated with the regulation of fatty acid metabolic process, instead of two if the semantic inheritance is ignored.

GO2PUB retrieves all gene names annotated by each GO term, directly or indirectly through the semantic inheritance properties. It then builds a query on the model "(*n* gene names, symbols or synonyms separated by OR) AND (*m* species) AND (*p* MeSH terms)". The name, symbol and synonyms of each gene compose the first part of the query. They will be searched in title and abstract. Species and keywords chosen by the users make up the second part of the query. Finally, GO2PUB submits to PubMed a query composed of gene names annotated directly or indirectly by the GO terms chosen by the users (name OR symbol OR Synonym), at least one species and some MeSH terms and free keywords. This big query is split into several smaller ones if it exceeds PubMed server URL length limitation. GO2PUB compiles the results and displays all citations numbered and sorted by date.

Endnotes

[a]www.ncbi.nlm.nih.gov/pubmed
[b]www.nlm.nih.gov/bsd/licensee/baselinestats.html
[c]http://www.geneontology.org/GO.ontology.relations.shtml
[d]go_daily-termdb-tables.tar.gz from http://archive.geneontology.org/latest-termdb/go_daily-termdb-tables.tar.gz
[e]ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/
[f]Arabidopsis, Chicken, Cow, Human, Mouse, Rat and Zebrafish
[g]http://www.ncbi.nlm.nih.gov/Taxonomy
[h]http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

Additional files

Additional file 1: This text file contents the experts selections for the query Q1.

Additional file 2: This text file contents the experts selections for the query Q2.

Additional file 3: This text file contents the experts selections for the query Q3.

Additional file 4: GO2PUB results file for query Q1.

Competing interest

We declare having no competing interest.

Author's contributions

CB developed GO2PUB method, software and website, participated in the design and the realization of the relevance study and drafted the manuscript. CD participated in the design and the realization of the relevance study and drafted the manuscript. AB participated in the design and the realization of the relevance study and drafted the manuscript. OD participated in the design and the realization of the relevance study and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Thanks to Biogenouest for computer support and hosting.
We are grateful to JF Ethier for his proofreading work.
CB was supported by a fellowship from the French ministry of research.

Received: 21 December 2011 Accepted: 26 August 2012
Published: 7 September 2012

References

1. Yoo S, Choi J: **On the query reformulation technique for effective MEDLINE document retrieval.** *J Biomed Inform* 2010, **43**(5):686–693.
2. Griffon N, Chebil W, Rollin L, Kerdelhue G, Thirion B, Gehanno JF, Darmoni SJ: **Performance evaluation of unified medical language system®'s synonyms expansion to query PubMed.** *BMC Med Inform Decis Mak* 2012, **12:**12.
3. Lu Z, Kim W, Wilbur WJ: **Evaluation of Query Expansion Using MeSH in PubMed.** *Inf Retr Boston* 2009, **12**(1):69–80.
4. Lu Z: **PubMed and beyond: a survey of web tools for searching biomedical literature.** *Database (Oxford)* 2011, **2011:**baq036.
5. Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, Stoehr P: **EBIMed–text crunching to gather facts for proteins from Medline.** *Bioinformatics* 2007, **23**(2):e237—e244.
6. Yamamoto Y, Takagi T: **Biomedical knowledge navigation by literature clustering.** *J Biomed Inform* 2007, **40**(2):114–130.
7. Tsuruoka Y, Miwa M, Hamamoto K, Tsujii J, Ananiadou S: **Discovering and visualizing indirect associations between biomedical concepts.** *Bioinformatics* 2011, **27**(13):i111–i119.
8. Bajpai A, Davuluri S, Haridas H, Kasliwal G, Deepti H, Sreelakshmi KS, Chandrashekar DS, Bora P, Farouk M, Chitturi N, Samudyata V, ArunNehru KP, Acharya K: **In search of the right literature search engine(s)** 2011. http://dx.doi.org/10.1038/npre.2011.2101.3.
9. Muin M, Fontelo P, Liu F, Ackerman M: **SLIM: an alternative Web interface for MEDLINE/PubMed searches - a preliminary study.** *BMC Med Inform Decis Mak* 2005, **5:**37.
10. Vanteru BC, Shaik JS, Yeasin M: **Semantically linking and browsing PubMed abstracts with gene ontology.** *BMC Genomics* 2008, **9** (Suppl 1):S10.
11. Bard JB, Rhee SY: **Ontologies in biology: design, applications and future challenges.** *Nat Rev Genet* 2004, **5:**213–222.
12. Bhogal J, Macfarlane A, Smith P: **A review of ontology-based query expansion.** *Inf Process & Manage* 2007, **43**(4):866–886.
13. Navigli R, Velardi P: **An analysis of ontology-based query expansion strategies.** In *Proceedings of the 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining,* Cavtat-Dubrovnik, Croatia; 2003:42–49 .

14. Matos S, Arrais JP, Maia-Rodrigues J, Oliveira JL: **Concept-based query expansion for retrieving gene related publications from MEDLINE.** *BMC Bioinformatics* 2010, **11:**212.
15. Harris MA, Consortium GO: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, **1:**D258–D261.
16. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R: **The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.** *Nucleic Acids Res* 2004, **32:**D262—D266.
17. Doms A, Schroeder M: **GoPubMed: exploring PubMed with the Gene Ontology.** *Nucleic Acids Res* 2005, **33:**W783–W786.
18. Rhee SY, Wood V, Dolinski K, Draghici S: **Use and misuse of the gene ontology annotations.** *Nat Rev Genet* 2008, **9:**509–515.
19. Cohen J: **A coefficient of agreement for nominal scales.** *Educational and Psychological Meas* 1960, **20**(1):37–46.
20. Landis JR, Koch GG: **The measurement of observer agreement for categorical data.** *Biometrics* 1977, **33**(1):159–174.
21. Hulsen T, de Vlieg J, Alkema W: **BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams.** *BMC Genomics* 2008, **9**(1):488.

doi:10.1186/2041-1480-3-7
Cite this article as: Bettembourg *et al.*: GO2PUB: Querying PubMed with semantic expansion of gene ontology terms. *Journal of Biomedical Semantics* 2012 **3**:7.

Conclusion

Notre outil GO2PUB a été capable de trouver des résultats pertinents ignorés par GoPubMed (Neuf résultats de la requête Q1 étaient spécifiques à GO2PUB, sept de la requête Q2 et trois de la requête Q3), même en procédant à une extension manuelle des requêtes de GoPubMed. À l'inverse, l'approche "*text-mining*" de GoPubMed a trouvé des articles pertinents ignorés par GO2PUB (Un résultat de la requête Q1 était spécifique à GoPubMed, trois de la requête Q2 et trois de la requête Q3). Cela démontre la pertinence de GO2PUB et sa complémentarité avec GoPubMed pour notre domaine d'intérêt. L'étude de généralisation a montré qu'un profil de résultat similaire était obtenu en utilisant des requêtes générées aléatoirement, en particulier lorsqu'on utilise des mots-clés pour préciser les requêtes. Cela suggère que les résultats de notre étude qualitative peuvent être généralisés.

Dans le cadre de cette thèse, GO2PUB a permi d'obtenir notamment de la bibliographie concernant le métabolisme des lipides chez l'Homme, la souris et la poule. Une grande partie des références présentes dans le chapitre 1 consacré au contexte biologique proviennent de GO2PUB.

2 Apport de la similarité sémantique dans la comparaison de gènes dupliqués

The Duplicated Genes Database : Identification and Functional Annotation of Co-Localised Duplicated Genes across Genomes

- Résumé -

Contexte : Le nombre d'études faisant un lien entre structure du génome et expression des gènes est en augmentation, avec une attention particulière portée aux gènes dupliqués. Bien qu'initialement dupliqués à partir d'une même séquence, les gènes dupliqués peuvent diverger fortement au cours de l'évolution et développer des fonctions ou une régulation d'expression différentes. Cependant, les informations concernant la fonction et l'expression des gènes dupliqués reste rare. Identifier des groupes de gènes dupliqués dans différent génomes et caractériser leur expression et leurs fonctions serait par conséquent d'un grand intérêt pour la communauté scientifique. La "*Duplicated Genes Database*" (DGD) a été développée dans ce but.

Méthodologie : DGD contient des données concernant neuf espèces. Pour chaque espèce, des analyses de BLAST ont été conduites sur les séquences peptidiques correspondant aux gènes localisés sur un même chromosome. Les groupes de gènes dupliqués ont été définis en fonction des résultats de leur comparaison via BLAST par paires et en fonction de la localisation chromosomique des gènes. Pour chaque groupe, les corrélations de Pearson entre les données d'expression des gènes ainsi que la similarité fonctionnelle basée sur les annotations GO de ces gènes ont été calculées.

Conclusion : La "*Duplicated Genes Database*" fournit une liste de gènes dupliqués colocalisés pour plusieurs espèces associée aux connaissances disponibles sur le niveau de co-expression et la valeur de similarité sémantique entre les gènes au sein de chaque groupe. L'ajout de ces données aux groupes de gènes dupliqués apporte une information biologique susceptible d'être utile à l'analyse d'expression de gènes. La "*Duplicated Genes Database*" est disponible sur le site de DGD à l'adresse `http://dgd.genouest.org`.

OPEN ACCESS Freely available online

PLOS ONE

The Duplicated Genes Database: Identification and Functional Annotation of Co-Localised Duplicated Genes across Genomes

Marion Ouedraogo[1,2☯], Charles Bettembourg[1,2☯], Anthony Bretaudeau[3], Olivier Sallou[3], Christian Diot[1,2], Olivier Demeure[1,2], Frédéric Lecerf[1,2]*

1 INRA, UMR1348 PEGASE, Saint-Gilles, France, **2** Agrocampus OUEST, UMR1348 PEGASE, Rennes, France, **3** GenOuest Platform, INRIA/Irisa – Campus de Beaulieu, Rennes, France

Abstract

Background: There has been a surge in studies linking genome structure and gene expression, with special focus on duplicated genes. Although initially duplicated from the same sequence, duplicated genes can diverge strongly over evolution and take on different functions or regulated expression. However, information on the function and expression of duplicated genes remains sparse. Identifying groups of duplicated genes in different genomes and characterizing their expression and function would therefore be of great interest to the research community. The 'Duplicated Genes Database' (DGD) was developed for this purpose.

Methodology: Nine species were included in the DGD. For each species, BLAST analyses were conducted on peptide sequences corresponding to the genes mapped on a same chromosome. Groups of duplicated genes were defined based on these pairwise BLAST comparisons and the genomic location of the genes. For each group, Pearson correlations between gene expression data and semantic similarities between functional GO annotations were also computed when the relevant information was available.

Conclusions: The Duplicated Gene Database provides a list of co-localised and duplicated genes for several species with the available gene co-expression level and semantic similarity value of functional annotation. Adding these data to the groups of duplicated genes provides biological information that can prove useful to gene expression analyses. The Duplicated Gene Database can be freely accessed through the DGD website at http://dgd.genouest.org.

Citation: Ouedraogo M, Bettembourg C, Bretaudeau A, Sallou O, Diot C, et al. (2012) The Duplicated Genes Database: Identification and Functional Annotation of Co-Localised Duplicated Genes across Genomes. PLoS ONE 7(11): e50653. doi:10.1371/journal.pone.0050653

Editor: Ramy K. Aziz, Cairo University, Egypt

Received April 28, 2012; **Accepted** October 24, 2012; **Published** November 28, 2012

Funding: This work was funded by INRA, Agrocampus Ouest and the Brittany Region. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: frederic.lecerf@agrocampus-ouest.fr

☯ These authors contributed equally to this work.

Introduction

A growing body of literature has shown that eukaryotic genomes contain groups of co-localised genes whose chromosomal location plays a role in the regulation of gene expression [1,2,3,4,5,6,7,8]. Part of these groups stems from gene duplications. Although duplicated genes are initially identical, they can evolve in different ways after the duplication event [9]. Some can remain co-regulated by retaining the same *cis*-regulatory motifs whereas others acquire different patterns of expression, resulting in un-correlated gene expression or even different tissue expression patterns. There may even be discrepancies in the co-expression patterns of duplicated genes depending on the genes or species analysed. In yeast [10] and *C. elegans* [11] for example, expression patterns are more similar between two duplicated genes than between two randomly-selected genes. Conversely, there are also reports of divergent profiles between duplicated genes according to expression level [12,13] and spatial expression [14,15,16,17,18]. Identifying groups of duplicated co-localised genes at a genomic scale for several species and characterizing both the expression and function of these genes would help bring a larger overview on this issue. While it is possible to get information on duplicated genes through a single gene query (i.e. Ensembl via its paralog genes list [19]), there is still no list of such duplicated genes available at genome-wide scale. Other tools dedicated to phylogeny studies only list duplicated genes without considering their co-location [20,21]. In addition, none of these tools give any information on gene expression level. Therefore, many researchers are forced to identify duplicated genes in their species of interest 'by hand' and then aggregate functional information from different sources [22,23,24,25,26,27,28,29,30].

This situation is further complexified by the fact that gene duplications can be divided into three major classes: 1) genomic-level duplications generated from whole genome or chromosomal duplication; 2) tandem duplications with genes closely localised in the same chromosome region; 3) other duplications corresponding

to genes with distant genomic locations [31]. In addition, recent studies also show that chromatin structures play a role in the co-expression of genes (for review, see [32]), including chromatin loops [33] or chromosome pairing in RNA factories [34,35]. Therefore, the co-location of genes may play a role in the regulation of their expression. For these reasons, we focused on tandem duplicated genes or groups of genes from multigene families (the above class 2 duplicated genes) further referred to as "groups of duplicated genes".

Here, we identified duplicated and co-localised genes from 9 different species. Co-expression and functional similarities between these duplicated genes were also determined. All this data is available through the Duplicated Genes Database (DGD) developed by our team.

Results

Database Implementation

The DGD workflow is depicted in Figure 1. In step one of the process, pairwise BLAST analyses were performed for each gene and each chromosome. These BLAST results were used with the genomic location of the genes to determine groups of co-localised duplicated genes. Gene annotations, i.e. name and description, were also added.

In step two of the process, gene co-expression and semantic similarity of GO annotations were determined. First, GEO expression data and GO annotations were retrieved for each duplicated gene. Then, after filtering the gene expression data, pairwise Pearson correlations were computed for each pair of genes in a group for each GEO dataset. The semantic similarity value for each pair was computed using the method of Wang [36].

The DGD website outputs this data in a dynamic image linking each gene in a group to the different values available.

Database Content

In total, the DGD contains 8411 groups of duplicated genes. By species, the number of groups varies from 444 in *Gallus gallus* (GGA) to 1412 in *Danio rerio* (DER) (Table 1). The number of duplicated genes also varies according to species, ranging from 1251 genes in GGA to 6036 in *Mus musculus* (MMU). Surprisingly, the majority of between-species variation comes from groups of 2 and 3 genes, whereas the numbers of groups of 4 and more genes are fairly similar (Figure 2). Mammalian species have similar patterns, except in *Sus scrofa* (SSC). The highest number of groups of 2 and 3 duplicated genes are found in DER (1132 groups) and SSC (1080 groups), while GGA has fewer duplicated groups than other species.

There are also differences between species according to size of the groups. The median size of duplicated groups is 105 kb in humans (HSA), with other species having fairly similar values, ranging from 58 kb in GGA to 248 kb in horse (ECA) (Table 2). Mean size is 641 kb in humans, and ranges from 601 kb in pig (SSC) to 1360 kb in rat (RNO). Gene number of the largest group is 77 in humans (corresponding to a group of olfactory receptor genes), and ranges from 428 genes in *Danio rerio* (corresponding to a Zinc finger genes group) down to 62 genes in *Gallus gallus* (an unidentified genes group as no annotations were available, although the Pfam database [37] reported a keratin domain).

The gap between species gets even larger when considering functional annotations and gene expression information. The percentage of groups of genes used for gene expression comparisons fluctuates strongly between humans (94%) or mice (93%) and fish (24%) or horse (0%). Similar variations exist for functional annotations: 83% and 88% of duplicated genes in humans and mice are annotated by GO terms in the GOA database *versus* just 12% and 25% in chicken and pig groups (Table 1).

Database Content Analyses

The pairwise Pearson correlations on the gene expression and semantic similarity values of the groups of duplicated genes were characterised in humans (Figures 3 and 4) and compared to results obtained from non-duplicated co-localised genes or randomly selected genes. These gene expression analyses were led on groups of 5 or less genes, as expression data for larger groups is often too incomplete to enable meaningful analysis. The same approach was applied for the analysis of semantic similarities in GO annotations (GOA), but with a maximum of 15 genes per group. Interestingly, the proportion of significant correlation was higher in groups of duplicated genes than in co-localised non-duplicated genes or genes randomly selected on the genome (figure 3A). The same results were observed when analyses were performed according to size of the group (figure 3B). Note that the proportion of significant correlation is similar between co-localised non-duplicated genes and genes randomly selected on the genome. Similar results were observed on semantic similarities, with higher values for duplicated genes than for randomly-selected genes whatever the number of genes in the group (figure 4A and 4B). This was not only the result of a higher proportion of electronic annotations (IEA) inferred from sequence similarities between these duplicated genes. Indeed, although IEA proportion increased with the number of duplicated genes in the groups, it was far lower in humans, for which 76% of the groups have been annotated, and in mouse, which is another 'well-annotated' species (88%), than in relatively 'poorly-annotated' species' such as ECA (42%) and SSC (at just 25%; see Table S1) in which most of the annotations are IEA (figure S1).

Database Interface

DGD has a web GUI handling queries in two major sections _ the browse page and the search page. The browse page gives direct access to database content for a species, a specific chromosome, or a defined genomic region. The search page allows users to run database queries for different terms using specific gene ID (Ensembl, Uniprot, RefSeq, GenBank, among others…), chromosomal location (chr:start.end) or any keywords (e.g. GTPase, death, fatty acids, etc.) that are searched for in the gene description. Users can perform multiple queries by typing several of these terms into the input box or by uploading a text file with the terms to search. In all cases, the search can be performed across all species or limited to a specific species. The DGD website search engine runs the query in the whole Ensembl dataset and cross-references database, and displays all the results even if the genes are not included in any co-localised and duplicated groups.

When a specific group of duplicated genes is selected, each gene is described by name (HGNC), by chromosome and by base pair location. The proportion of experiments with significant correlation of expression and the semantic similarities between genes in biological process, molecular function and cellular component gene ontology terms are also shown as a graph if the information is available.

Cross-references can be added to this display (functional annotation, various gene IDs from others databases). Users should note that the lists of cross-references are species-dependent, and so this feature is disabled for queries across all the species. The display gives hyperlinks to the selected cross-reference databases.

For both browse pages and search pages, users can choose between different export formats or display modes (lists of genes or lists of groups, in tab-delimited file format).

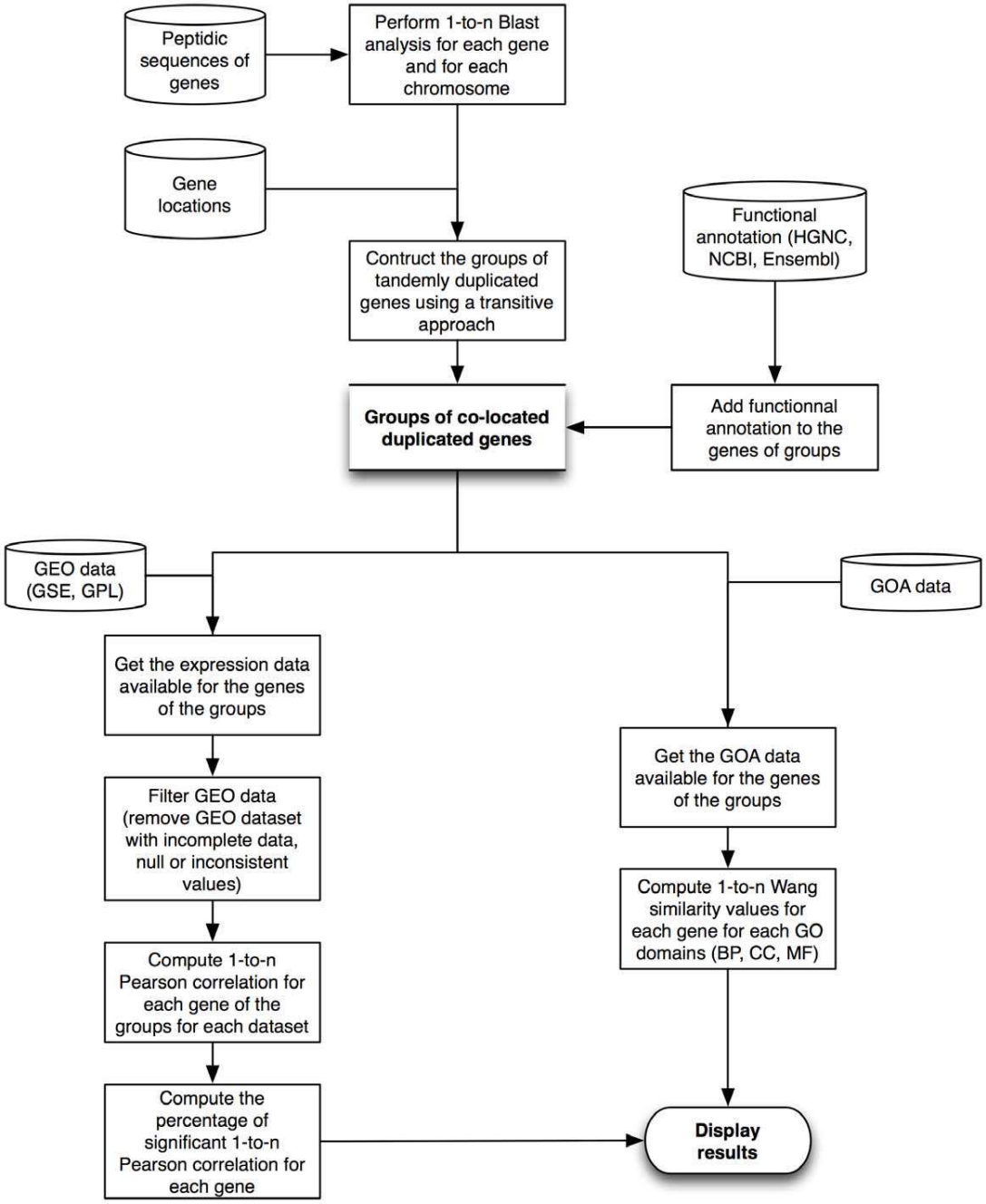


Figure 1. DGD workflow. Description of the DGD database development process, from sequence similarity analyses and integration of gene annotation data from NCBI, Ensembl and HGNC websites to the integration and computation of functional data from GEO (Gene Expression Omnibus) and GOA (Gene Ontology Annotation).
doi:10.1371/journal.pone.0050653.g001

DGD is publicly available as a SOAP web service that has been implemented in Java using the Opal2 toolkit [38]. The DGD web service only accepts Ensembl gene IDs as search input and cannot return external references directly. However, a second web service named Xref dedicated to cross-references management is available on the Genouest server [39]. For a given set of genes, the Xref web service searches corresponding Ensembl genes using cross-references, and returns a set of external references for the given set of genes. Thus, users should use the Xref web service in contexts when they need conversions between Ensembl gene IDs and other identifiers. Full developer documentation, WSDL files, code examples, and Taverna workflows are all available for both services via the DGD website.

Discussion

The goal of the DGD database was to provide information on co-localised duplicated genes. To this end, two parameters had to be defined: the sequence similarity threshold between two genes,

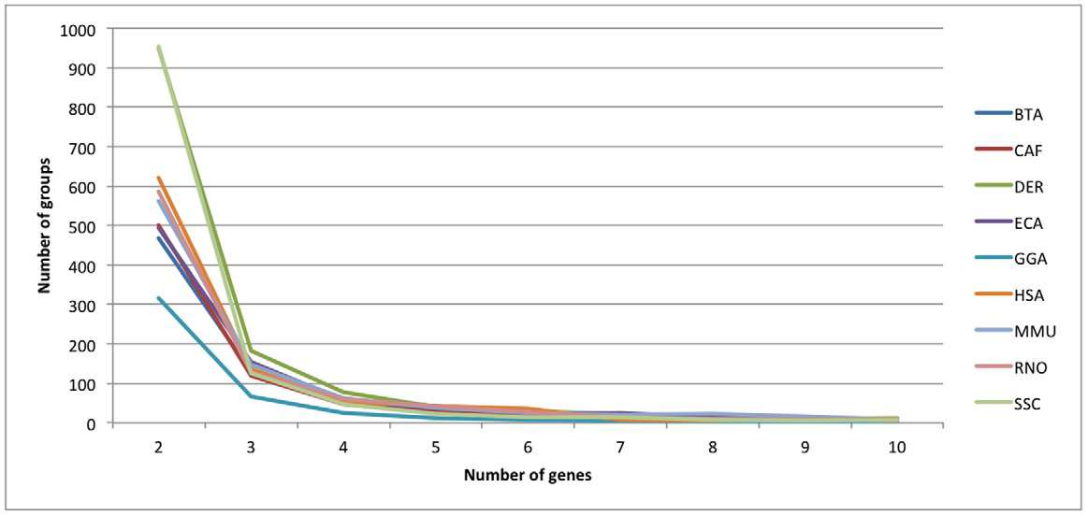


Figure 2. Distribution of the number of groups of duplicated genes according to number of duplicated genes. BTA: *Bos taurus*; CAF: *Canis familiaris*; DER: *Danio rerio*; ECA: *Equus caballus*; GGA: *Gallus gallus*; HSA: *Homo sapiens;* MMU: *Mus musculus*; RNO: *Rattus norvegicus* and SSC: *Sus scrofa*.
doi:10.1371/journal.pone.0050653.g002

and the maximum distance defining duplicated genes as co-localised. The literature features various different approaches developed for detecting duplicated genes. Most of these approaches revolve around sequence comparisons using either FASTA [9,16,40] or BLAST [28,41,42]. The threshold values defined by these comparison tools are generally based on 1) a first selection based on an e-value threshold to remove non-relevant sequence comparison results, and 2) the value defined by Rost [43], who proposed a formula using percentage identity and length of the alignment between the two sequences. Note that some studies have only used the e-value and a minimum alignment coverage threshold [25,42]. Here, we applied another approach first proposed by Li *et al.* [44] that computes another identity value *I'*, weighting the initial identity value with the number of amino acids and the length of the aligned region. This improvement avoids the clustering of non-homologous genes that share the same domain, such as when a short protein shares domains with a longer protein. The threshold values proposed by Li *et al.* were used to define the groups of pairwise duplicated genes (i.e. I'≥30% for alignment >150 aa and I'≥p' from Rost for alignment <150 aa). Using these more stringent thresholds instead of those of the Ensembl database (2%–24%) results in a conservative approach that is expected to reduce the number of false-positives.

Another major parameter that dictates the definition of groups of duplicated genes is size of the gene window. In the literature, the maximum distance within which duplicated genes are considered as co-localised is defined using either a physical distance [22,27] or a window including *n* genes [29,30]. The physical distance approach may be more stringent but it has a major pitfall: as genome length and gene density are not the same in the different species, the distance has to be defined in a species-specific way (from 200 kb for *C. elegans* to 1 Mb for *H. sapiens*, for instance). The gene window approach, however, is compatible with many species and is not sensitive to gene density variability between chromosomes and between species. Here, duplications were searched within a window of 100 genes. Although at first sight this may seem a large number, the median size of the duplicated groups reported here was 105 kb in humans and was fairly similar in other species, with values ranging from 58 kb in chicken to 248 kb in horse. This suggests that the duplicated genes identified are closely localised, and that defining distance as a number of genes rather than a physical distance does not greatly affect the genomic size of the groups.

The total number of groups of duplicated genes differs between species (Figure 2). These differences are observed mainly in groups containing two or three duplicated genes and between mammalian species and other species. In mammals, the only exception is the pig, for which the genome assembly is of poor quality, which could lead to the identification of false-positive groups of duplicated genes. This artificially increases the number of small groups of

Table 1. Statistics on DGD content.

	HSA	MMU	RNO	CAF	GGA	BTA	DER	ECA	SSC
Total peptides	74640	40732	32948	25559	22194	26977	28630	22641	19083
Non-redundant peptides	47313	30659	24812	22383	19371	23833	26204	21551	18273
Groups	964	1008	959	751	444	798	1412	894	1229
Genes in groups	3710	6036	4899	2647	1251	3714	5830	4601	4210

For each species (*Bos taurus* (BTA), *Danio rerio* (DER), *Canis familiaris* (CAF), *Gallus gallus* (GGA), *Equus caballus* (ECA), *Homo sapiens* (HSA), *Mus musculus* (MMU), *Rattus norvegicus* (RNO) and *Sus scrofa* (SSC)), the numbers of peptide sequences used in the analyses (only non-redundant) are reorted here with the number of peptide sequences initially available (total).
doi:10.1371/journal.pone.0050653.t001

Table 2. Statistics for the groups of duplicated genes.

	HSA	MMU	RNO	CAF	GGA	BTA	DER	ECA	SSC
Mean group size (kb)	641	1007	1360	1317	892	1167	666	3368	601
Median group size (kb)	105	144	235	165	58	154	111	248	151
Maximum number of genes in largest groups	77	267	217	133	62	174	428	171	164

For each species (*Bos taurus* (BTA), *Danio rerio* (DER), *Canis familiaris* (CAF), *Gallus gallus* (GGA), *Equus caballus* (ECA), *Homo sapiens* (HSA), *Mus musculus* (MMU), *Rattus norvegicus* (RNO) and *Sus scrofa* (SSC)), the mean and median genomic size (in kb) of the groups and the maximum number of genes in the largest groups are indicated.
doi:10.1371/journal.pone.0050653.t002

duplicated genes. In chicken and zebrafish, part of the differences could be assigned to the phylogeny distance with mammals [45].

Every species featured some very large groups, ranging from 62 genes in GGA to 428 genes in DER. In humans, the largest groups include T-cell receptor genes, zing finger genes, immunoglobulin genes, or notoriously highly duplicated olfactory receptor genes [46]. In fact, it is possible to find clear false-positive groups due to errors in the genome assemblies, especially for most current genomes that, like the pig, are what Yandel and Ence (2012) called 'standard draft assembly' genomes [47]. However, as the DGD database is updated at each Ensembl update cycle, we expect to see genome assembly errors fixed in the future.

Gene co-expression level and functional similarity in GO annotations can be combined inside a group by computational processes on data from GEO and GOA. We thus tested a few hypotheses using the human data. The first and highly controversial hypothesis is that gene co-expression might be higher in groups of duplicated genes than in groups of randomly-selected genes [10,11,12,13]. As illustrated in Figure 3A, co-localised duplicated genes have a higher proportion of significant co-expression than co-localised non-duplicated genes or genes randomly selected in the genome. This difference is observed whatever the number of genes within the groups (Figure 3B).

Another interesting hypothesis to test was whether there is functional conservation or divergence between duplicated genes [9]. Comparing GO semantic similarities between co-localised duplicated genes against randomly-selected genes revealed that annotated biological processes present much higher similarities between co-localised duplicated genes (Figure 4A). Surprisingly, the similarity between genes significantly increases with group size (Figure 4B). This is probably due to a lack of "specific" annotation when the number of duplicated genes does not allow experimental validations. Indeed, for most of the genes annotated in the large duplicated groups, the annotation was automatically inferred from electronic annotation (IEA evidence code). As shown in figure S1, this is particularly true in species for which annotation is qualified as "poor quality", the best examples being ECA and SSC with 42% and 25%, respectively, of the groups annotated with almost all GO terms inferred electronically (IEA), but less so in model species (HSA, MMU, and to a lesser extent RNO) for which annotation is qualified as "good quality". Taken together, these results clearly suggest that, at least in humans, tandem and multi-duplicated genes show higher co-expression levels and similarity of functional GO annotations than other genes.

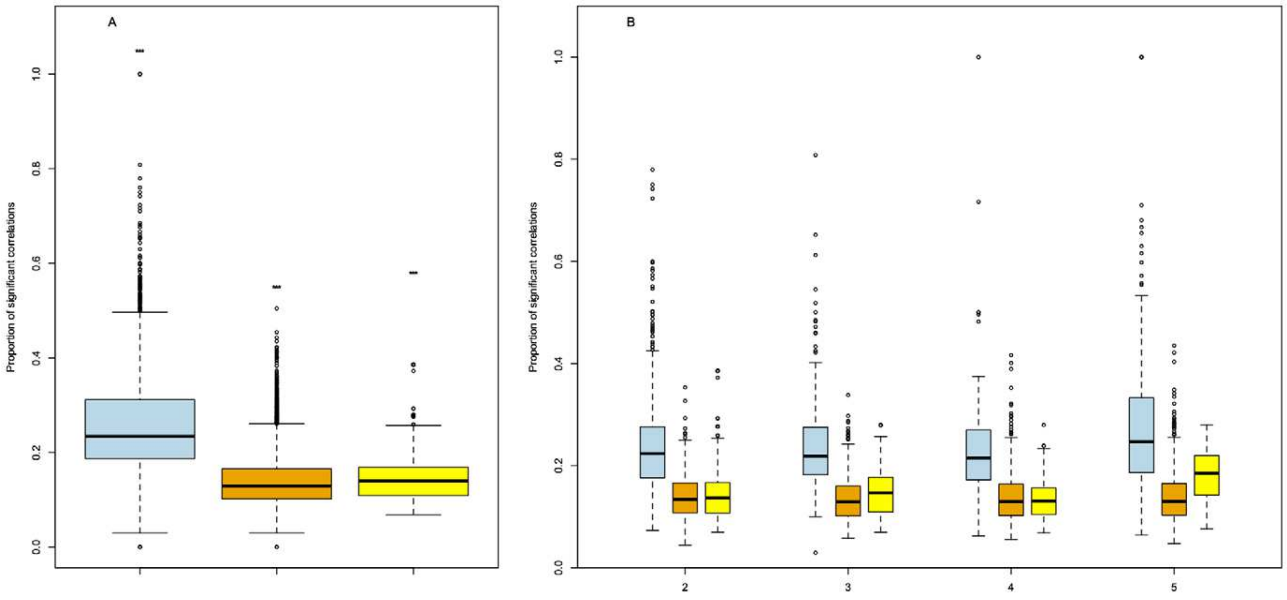


Figure 3. Proportion of significant correlations. Boxplots of significant correlations of expression for duplicated genes (blue), non-duplicated genes (orange) and randomly-selected genes (yellow). (**A**) Correlations for all groups of genes. Means with a different letter are significantly different according to Student's R t-tests at $p<0.05$ (n = 3320, 2760 and 13605, respectively). (**B**) Correlations according to the number of genes within groups. For every group size, the means of each type of group are significantly different ($p<0.05$).
doi:10.1371/journal.pone.0050653.g003

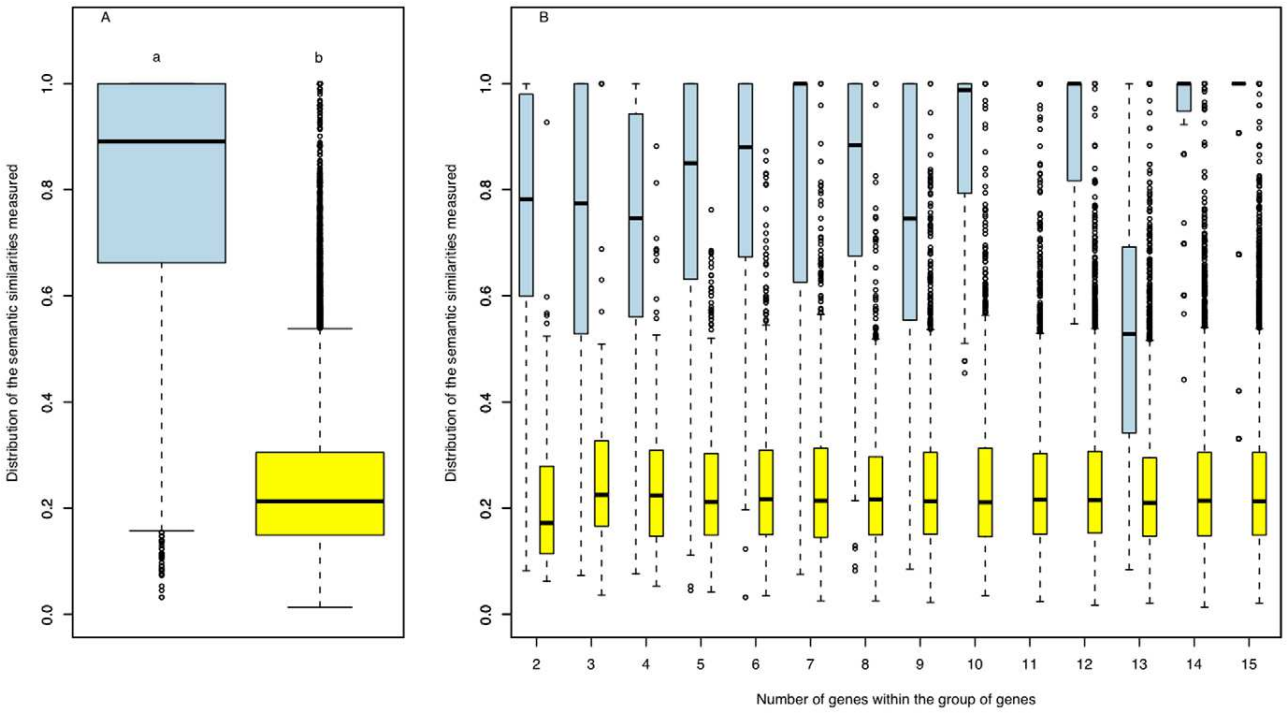


Figure 4. Distribution of semantic similarities. (**A**) Distribution of GO biological process semantic similarities in duplicated gene groups (blue) *vs.* randomly-selected gene groups (yellow). Means with a different letter are significantly different according to Student's R t-tests at $p<0.05$. (**B**) Details of the same distribution with groups pooled by size. The mean of each duplicated group is significantly different from the mean of each randomly-selected genes group ($p<0.05$). *Note: no data were available for the group with 11 genes.*
doi:10.1371/journal.pone.0050653.g004

Conclusion

This database provides a simple way to quickly and easily find groups of tandem duplicates or large groups of multigene families by gene identifier, chromosomal location and/or keywords. Gene co-expression level and semantic similarities in functional annotations are also displayed when raw data is available. DGD is the first database to integrate this genomic information on co-localised duplicated genes with gene expression data and GO annotation similarity. This database can be readily expanded to other genomes as long as genomic annotations and peptide sequences are available.

Materials and Methods

Sequence Data

As shown in Figure 1, peptide sequences and chromosomal location of the genes were downloaded from the Ensembl FTP site [48] (Ensembl version 68) for 9 species: *Bos taurus* (BTA), *Danio rerio* (DER), *Canis familiaris* (CAF), *Gallus gallus* (GGA), *Equus caballus* (ECA), *Homo sapiens* (HSA), *Mus musculus* (MMU), *Rattus norvegicus* (RNO) and *Sus scrofa* (SSC). For each gene, only the longest peptide sequence was kept (peptide sequence numbers are given in Table 1).

Identification of Duplicated Genes

Duplicated genes were identified using a two-step strategy. For each genome, a BLAST search was conducted between all peptide sequences of the genes in a chromosome. To determine whether two peptides were similar, we computed identity $I' = I \times Min(n_1/L_1, n_2/L_2)$ proposed by Li *et al.* [44], where I is the proportion of identical amino acids in the aligned region (including gaps) between sequences 1 and 2, L_i is the length of sequence i, and n_i is the number of amino acids in the aligned region in sequence i. Two genes were considered duplicates if an all-against-all BLAST search within a window of 100 genes [29,30] met the following criteria: i) e-value is ≤ 0.2 (only to filter non-relevant BLAST results); ii) I' $\geq 30\%$ if L ≥ 150 a.a. (where L is the length of the aligned region) or I $\geq 0.01n+4.8L^{-0.32(1+\exp(-L/1000))}$ [43] if L <150 a.a. (where $n=6$ as it makes the formula continuous at $L=150$), as proposed by Li *et al.* [44]. Within the best BLAST hits for a given gene query, we selected the "hit" gene that had the closest chromosomal location downstream of the gene queried.

Duplicated gene groups were then put together based on the principle of a simple transitive link between the remaining genes: if gene A was similar to gene B and to gene C, then genes A, B and C were included in the same group, even if genes B and C were not found similar. Chromosomal location information and gene annotations (name and description) of each gene for all duplicated groups were then incorporated into a MySQL database.

Database Objects

For each species, Ensembl cross-references [48] were integrated into the MySQL database to enable queries on specific genes using an Ensembl or HGNC keyword. In addition, data on Ensembl objects (genes, transcripts and translations) as well as other database objects (NCBI, etc.) were also collected to be displayable in the results page if needed. The list of available reference sources was specific to each species depending on the sources found in the Ensembl dataset. For each gene, the external references displayed are those associated to the gene and to any of its transcripts and any of the corresponding translations.

Functional gene annotations were retrieved from the Gene Ontology Annotation (GOA) database [49]. The GO structure used to compute similarity was obtained from the term and term2term tables of the GO database [50].

All database updating procedures have been incorporated into the BioMaj workflow engine [51] to integrate future updates at each new Ensembl database version.

Gene Expression Correlations Using GEO

The HGNC id of each duplicated gene was searched through the annotation platform (GPL) of the Gene Expression Omnibus (GEO) database [52]. The corresponding GEO experiments (GSE) were extracted. Only GSE expression data that satisfied the following conditions were kept: a) a minimal number of 3 samples available; b) the genes of a duplicated group were all present within the GSE; c) GSE with null values or always the same value were discarded.

For each group of duplicated genes and for each GSE, the Pearson correlation and associated *p*-value were computed between each gene pair using a bilateral test, and the proportion of significant correlations for each gene pair within a group of duplicated genes was retrieved.

To assess whether co-localised duplicated genes had a higher proportion of significant correlations, we ran this same procedure on non-duplicated genes that were selected as i) co-localised or ii) randomly distributed among the human genome. The proportions of significant correlations between conditions were tested using a *Student* t-test.

Similarities in GO Annotations

Semantic similarities in GO annotations were determined using Wang's method [36] and computed pairwise in a group every time at least two annotated genes were found. As GO is split into three different branches – Biological Process, Molecular Function and Cellular Component – three similarity values were computed for each pairwise comparison. All the similarity values calculated with this method were bounded from 0 to 1. The higher the similarity value, the more the compared genes shared the same biological functions. Wang considers two genes as fairly similar at a similarity value of 0.5.

Supporting Information

Figure S1 Proportion of IEA according to duplicated gene number in the groups in nine species.
(TIF)

Table S1 Description of DGD groups annotated for Gene Ontology. For each species, the number of groups, the number of annotated groups with GO terms and the percentage of groups annotated are indicated.
(DOC)

Acknowledgments

The authors would like to thank the GenOuest platform for the hosting the DGD. The authors thank A.T.T scientific editing services for proofreading the manuscript.

Author Contributions

Conceived and designed the experiments: FL OD. Performed the experiments: FL CB MO. Analyzed the data: FL CB MO CD OD. Wrote the paper: FL OD AB CD. Integration of DGD in the GenOuest dataframe: OS AB.

References

1. Barrans JD, Ip J, Lam C-W, Hwang IL, Dzau VJ, et al. (2003) Chromosomal distribution of the human cardiovascular transcriptome. Genomics 81: 519–524.
2. Bortoluzzi S, Rampoldi L, Simionati B, Zimbello R, Barbon A, et al. (1998) A comprehensive, high-resolution genomic transcript map of human skeletal muscle. Genome Res 8: 817–825.
3. Ko MS, Threat TA, Wang X, Horton JH, Cui Y, et al. (1998) Genome-wide mapping of unselected transcripts from extraembryonic tissue of 7.5-day mouse embryos reveals enrichment in the t-complex and under-representation on the X chromosome. Hum Mol Genet 7: 1967–1978.
4. Minagawa S, Nakabayashi K, Fujii M, Scherer SW, Ayusawa D (2004) Functional and chromosomal clustering of genes responsive to 5-bromodeoxyuridine in human cells. Experimental Gerontology 39: 1069–1078.
5. Purmann A, Toedling J, Schueler M, Carninci P, Lehrach H, et al. (2007) Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality. Genomics 89: 580–587.
6. Soury E, Olivier E, Simon D, Ruminy P, Kitada K, et al. (2001) Chromosomal assignments of mammalian genes with an acute inflammation-regulated expression in liver. Immunogenetics 53: 634–642.
7. Vogel JH, von Heydebreck A, Purmann A, Sperling S (2005) Chromosomal clustering of a human transcriptome reveals regulatory background. BMC Bioinformatics 6: 230.
8. Zhang H, Pan K-H, Cohen SN (2003) Senescence-specific gene expression fingerprints reveal cell-type-dependent physical clustering of up-regulated chromosomal loci. Proceedings of the National Academy of Sciences of the United States of America 100: 3251–3256.
9. Zhang P, Gu Z, Li W-H (2003) Different evolutionary patterns between young duplicate genes in the human genome. Genome Biology 4: R56–R56.
10. Zhang Z, Gu J, Gu X (2004) How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? Trends Genet 20: 403–407.
11. Castillo-Davis CI, Hartl DL, Achaz G (2004) cis-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res 14: 1530–1536.
12. Gu Z, Rifkin SA, White KP, Li W-H (2004) Duplicate genes increase gene expression diversity within and between species. Nat Genet 36: 577–579.
13. Huminiecki L, Wolfe KH (2004) Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res 14: 1870–1879.
14. Blanc G, Wolfe KH (2004) Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution. Plant Cell 16: 1679–1691.
15. Ganko EW, Meyers BC, Vision TJ (2007) Divergence in Expression between Duplicated Genes in Arabidopsis. Mol Biol Evol 24: 2298–2309.
16. Gu Z, Nicolae D, Lu HHS, Li WH (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. Trends in Genetics: TIG 18: 609–613.
17. Li W-H, Yang J, Gu X (2005) Expression divergence between duplicate genes. Trends in Genetics: TIG 21: 602–607.
18. Makova KD, Li W-H (2003) Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Research 13: 1638–1645.
19. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, et al. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Research 19: 327–335.
20. Duret L, Perriere G, Gouy M (1999) "HOVERGEN: database and software for comparative analysis of homologous vertebrate genes". In: Letovsky S, editor. Bioinformatics Databases and Systems. Boston: Kluwer Academic Publishers. 13–29.
21. Van de Peer Y, Taylor JS, Joseph J, Meyer A (2002) Wanda: a database of duplicated fish genes. Nucleic Acids Res 30: 109–112.
22. Lercher MJ, Blumenthal T, Hurst LD (2003) Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. Genome Res 13: 238–243.
23. Farré D, Albà MM (2010) Heterogeneous patterns of gene-expression diversification in mammalian gene duplicates. Molecular Biology and Evolution 27: 325–335.
24. Chung W-Y, Albert R, Albert I, Nekrutenko A, Makova K (2006) Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network. BMC Bioinformatics 7: 46–46.
25. Fukuoka Y, Inaoka H, Kohane IS (2004) Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. BMC Genomics 5: 4–4.
26. Ren X-Y, Fiers MWEJ, Stiekema WJ, Nap J-P (2005) Local Coexpression Domains of Two to Four Genes in the Genome of Arabidopsis. Plant Physiol 138: 923–934.
27. Lercher MJ, Urrutia AO, Hurst LD (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet 31: 180–183.

28. Li Q, Lee BT, Zhang L (2005) Genome-scale analysis of positional clustering of mouse testis-specific genes. BMC Genomics 6: 7.
29. Ng YK, Wu W, Zhang L (2009) Positive correlation between gene coexpression and positional clustering in the zebrafish genome. BMC Genomics 10: 42.
30. Williams EJ, Bowles DJ (2004) Coexpression of neighboring genes in the genome of Arabidopsis thaliana. Genome Res 14: 1060–1067.
31. Jianzhi Z (2003) Evolution by gene duplication: an update. Trends in Ecology & Evolution 18: 292–298.
32. Baker M (2011) Genomics: Genomes in three dimensions. Nature 470: 289–294.
33. Kadauke S, Blobel GA (2009) Chromatin loops in gene regulation. BBA - Gene Regulatory Mechanisms 1789: 17–25.
34. Xu M, Cook PR (2008) The role of specialized transcription factories in chromosome pairing. Biochimica Et Biophysica Acta 1783: 2155–2160.
35. Xu M, Cook PR (2008) Similar active genes cluster in specialized transcription factories. The Journal of Cell Biology 181: 615–623.
36. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF (2007) A new method to measure the semantic similarity of GO terms. Bioinformatics (Oxford, England) 23: 1274–1281.
37. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein families database. Nucleic Acids Research 40: D290–301.
38. Krishnan S, Clementi L, Ren J, Papadopoulos P, Li W. Design and Evaluation of Opal2: A Toolkit for Scientific Software as a Service; 2009; Los Alamitos, CA, USA. IEEE Computer Society. 709–716.
39. Genouest (2012) Genouest Xref server: a webservice dedicated to cross-references.
40. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, et al. (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421: 63–66.
41. Friedman R, Hughes AL (2001) Gene duplication and the structure of eukaryotic genomes. Genome Research 11: 373–381.
42. Hsiao T-L, Vitkup D (2008) Role of Duplicate Genes in Robustness against Deleterious Human Mutations. PLoS Genet 4: e1000014–e1000014.
43. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12: 85–94.
44. Li WH, Gu Z, Wang H, Nekrutenko A (2001) Evolutionary analyses of the human genome. Nature 409: 847–849.
45. Hedges SB (2002) The origin and evolution of model organisms. Nat Rev Genet 3: 838–849.
46. Niimura Y, Nei M (2006) Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates. Journal of human genetics 51: 505–517.
47. Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13: 329–342.
48. The Ensembl FTP Server: ftp.ensembl.org/pub/current_fasta/.
49. Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, et al. (2012) The UniProt-GO Annotation database in 2011. Nucleic Acids Research 40: D565–570.
50. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25: 25–29.
51. Filangi O, Beausse Y, Assi A, Legrand L, Larre JM, et al. (2008) BioMAJ: a flexible framework for databanks synchronization and processing. Bioinformatics 24: 1823–1825.
52. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, et al. (2011) NCBI GEO: archive for functional genomics data sets–10 years on. Nucleic Acids Research 39: D1005–1010.

Conclusion

La "*Duplicated Genes Database*" (DGD) procure une moyen simple et rapide de trouver des groupes de gènes dupliqués en tandem ou des grands groupes composés de familles multi-géniques à partir d'un identifiant de gène, d'une localisation chromosomique ou de mots-clés. Les données de co-expression des gènes ainsi que les valeurs de similarités de leurs annotations fonctionnelles (GO) sont aussi fournies quand disponibles. DGD est la première base de données à intégrer cette information génomique sur des gènes dupliqués co-localisés avec leurs données d'expression et les similarités basées sur leurs annotations GO. Cette base de données peut être aisément étendue à d'autres génomes tant que des annotations génomiques et des séquences peptidiques sont disponibles.

Ces travaux ont été réalisés avant le développement de notre mesure de particularité sémantique, qu'il serait intéressant d'appliquer aux gènes des groupes de DGD. Il est également intéressant de noter que d'après le boxplot de la Figure 4 qui concerne la comparaison sur BP d'ensembles de gènes a priori similaires (S) car dupliqués et non similaires (N), le seuil de similarité τ_{sim} de 0.4 que nous avons obtenu un an après les travaux présentés ici tombe entre la moustache inférieure de la boîte bleue (S) et la moustache supérieure de la boîte jaune (N), désignées respectivement par τ_N et τ_S dans l'article du chapitre 4. Les distributions utilisées ici se comportent comme celles de cet article et confortent la démarche utilisée pour définir un seuil de similarité.

3 Étude de l'évolution de la complexité de Gene Ontology

Measuring the Evolution of Ontology Complexity : The Gene Ontology Case Study

- Résumé -

Les ontologies supportent le partage automatique, la combinaison et l'analyse de données biologiques. Elles subissent une maintenance et un enrichissement régulier. Nous avons étudié l'influence de l'évolution d'une ontologie sur sa complexité structurale. Comme étude de cas, nous avons utilisé les soixante versions mensuelles entre janvier 2008 et décembre 2012 de Gene Ontology et de ses trois branches indépendantes, i.e. biological processes (BP), cellular components (CC) et molecular functions (MF). Pour chaque cas, nous avons mesuré la complexité en calculant des métriques relatives à la taille, la connectivité des nœuds et la structure hiérarchique. Le nombre de classes et de relations a augmenté de manière monotone pour chaque branche, avec différents taux de croissance. Nous avons constaté que BP et CC ont une connectivité similaire, supérieure à celle de MF. La connectivité a augmenté de façon monotone pour BP, diminué pour CC et est restée stable pour MF, avec une augmentation marquée pour les trois branches en novembre et décembre 2012. Les mesures relatives à la hiérarchie ont montré que CC et MF avaient des proportions similaires de feuilles, et une profondeur moyenne et hauteur moyenne similaires également. BP avait une plus faible proportion de feuilles ainsi qu'une profondeur moyenne et une hauteur moyenne plus haute. Pour BP et MF, l'augmentation de connectivité fin 2012 a résulté en une augmentation de la profondeur moyenne et de la hauteur moyenne et en une diminution de la proportion de feuilles, indiquant un effort d'enrichissement majeur du niveau moyen de la hiérarchie. La variation du nombre de classes et de relations dans une ontologie ne donne pas assez d'information au sujet de l'évolution de sa complexité. Cependant, la connectivité et les métriques relatives à la hiérarchie ont révélé des profils différent de valeurs aussi bien que d'évolution pour les trois branches de Gene Ontology. Nous avons constaté que CC était similaire à BP en terme de connectivité, et similaire à MF en terme de hiérarchie des concepts. Globalement, la complexité de BP a augmenté, CC a subi un ajout de feuilles permettant un niveau d'annotations plus fin, mais diminuant légèrement sa complexité, et la complexité de MF est restée stable.

OPEN ACCESS Freely available online

PLOS ONE

Measuring the Evolution of Ontology Complexity: The Gene Ontology Case Study

Olivier Dameron[1,2]*, Charles Bettembourg[1,2,3], Nolwenn Le Meur[1,4]

1 Université de Rennes, Rennes, France, **2** Institut de Recherche en Informatique et Systèmes Aléatoires, Rennes, France, **3** UMR1348 PEGASE Institut national de la recherche agronomique – Agrocampus OUEST, Rennes, France, **4** Ecole des hautes études en santé publique, Rennes, France

Abstract

Ontologies support automatic sharing, combination and analysis of life sciences data. They undergo regular curation and enrichment. We studied the impact of an ontology evolution on its structural complexity. As a case study we used the sixty monthly releases between January 2008 and December 2012 of the Gene Ontology and its three independent branches, i.e. biological processes (BP), cellular components (CC) and molecular functions (MF). For each case, we measured complexity by computing metrics related to the size, the nodes connectivity and the hierarchical structure. The number of classes and relations increased monotonously for each branch, with different growth rates. BP and CC had similar connectivity, superior to that of MF. Connectivity increased monotonously for BP, decreased for CC and remained stable for MF, with a marked increase for the three branches in November and December 2012. Hierarchy-related measures showed that CC and MF had similar proportions of leaves, average depths and average heights. BP had a lower proportion of leaves, and a higher average depth and average height. For BP and MF, the late 2012 increase of connectivity resulted in an increase of the average depth and average height and a decrease of the proportion of leaves, indicating that a major enrichment effort of the intermediate-level hierarchy occurred. The variation of the number of classes and relations in an ontology does not provide enough information about the evolution of its complexity. However, connectivity and hierarchy-related metrics revealed different patterns of values as well as of evolution for the three branches of the Gene Ontology. CC was similar to BP in terms of connectivity, and similar to MF in terms of hierarchy. Overall, BP complexity increased, CC was refined with the addition of leaves providing a finer level of annotations but decreasing slightly its complexity, and MF complexity remained stable.

Citation: Dameron O, Bettembourg C, Le Meur N (2013) Measuring the Evolution of Ontology Complexity: The Gene Ontology Case Study. PLoS ONE 8(10): e75993. doi:10.1371/journal.pone.0075993

Editor: Marc Robinson-Rechavi, University of Lausanne, Switzerland

Received June 27, 2013; **Accepted** August 20, 2013; **Published** October 11, 2013

Funding: No current external funding sources for this study.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: olivier.dameron@univ-rennes1.fr

Introduction

The problem of ontology quality variation

Ontologies are instrumental for sharing, combining and analyzing life sciences data [1]. Ontologies evolve through regular modifications related to curation or to enrichment [2]. Existing metrics quantifying the changes rely on the variation of the number of classes, of the number of properties, or for the most sophisticated, of the number of restrictions [3]. For example, the Ontology Evolution Explorer OnEX provides access to approximately 560 versions of 16 life science ontologies. It allows a systematic exploration of the changes by generating evolution trend charts and inspection of the added, deleted, fused and obsolete concepts [4]. The underlying assumption of these approaches is that for ontologies, the more classes and properties, the better.

However, the creation of a new class could decrease the overall quality of the ontology, whereas previous measures would increase. Likewise, deleting an erroneous class would increase the overall quality of the ontology, but previous measures would decrease. Moreover, these measures are not affected if one class is moved from one location to another, nor if one class is deleted and another one added.

Related general approaches

Together with OnEX, GOMMA is a generic infrastructure for managing and analyzing life science ontologies and their evolution [3]. It provides advanced comparison capabilities of two versions of an ontology. Its Region Analyzer identifies evolving and stable regions of ontologies by determining the cost of different change operations such as deletions and additions.

Malone and Stevens measured the activity of an ontology by analyzing the additions, deletions and changes as well as the regularity and frequency of releases [5] on 5036 versions of 43 ontologies. They successfully identified five profiles of activity (initial, expanding, refining, optimizing and dormant).

While the previous two approaches focused on changes by analyzing ontology variations, others took a static perspective on ontology analysis. OntoClean is a formal method for structuring and analyzing ontologies based on metaproperties of classes (identity, unity, rigidity and dependence) [6]. To our knowledge, there is no effort to apply this method to the GO. Kölher et al. developed the GULO (Getting an Understanding of LOgical definitions) Java package for automatic reasoning on classes logical definitions [7]. Its exploits the logical definitions and the explicit cross-references between ontologies to compare the relations in the ontology of interest with relations inferred from the references

ontologies. This facilitates the systematic detection of omissions and incompatibilities. Shchekotykhin et al. proposed an entropy-based approach for localizing faults when debugging ontologies [8]. Yao et al. formally defined metrics of an ontology's fit with respect to published knowledge in the form of other ontologies and of scientific articles [9]. Hoehndorf et al. propose a method to evaluate biomedical ontologies for a particular problem by quantifying the success of using the ontology for this problem [10]. Comparing the measures of success of two versions of an ontology for the same problem would provide an indication of the relevance of the modifications.

These generic solutions were completed by various ontology-specific efforts to detect inconsistencies or ambiguities, such as the Unified Medical Language System (UMLS) [11], the Medical Entities Dictionary [12], the Cancer Biomedical Informatics Grid (CaBIG) [13], the NCI Thesaurus (NCIt) [14]. Other approches relied on the ontology structure, e.g. for the Foundational Model of Anatomy (FMA) [15] or on logical definitions of classes, e.g. on the Cell Ontology [16] or SNOMED-CT [17].

Yao et al. provide a review of ontology evaluation and identified four categories: (1) measures of an ontology's internal consistency, (2) usability and task-based performance, (3) comparison with other ontologies and (4) match to reality [9].

Ontology complexity as a measurable proxy for ontology quality

There is a need for a finer grain measure of the quality of an ontology which would allow a better assessment of the impact of a change or of a set of changes. One of the difficulties of defining and measuring the quality of an ontology is that it refers to how well the ontology reflects reality, of which we have an incomplete and imperfect understanding. Ontology complexity is an aspect of quality more amenable to formal analysis. Moreover, it focuses on an intrinsic feature of an ontology, not its suitability for a particular task.

None of the previous general efforts addresses the question of the impact of the changes on the ontology complexity. We propose an approach based on ontology complexity. Compared to Yao et al.'s four categories of ontology evaluation [9], it offers a complementary view but is different from ontology's internal consistency.

Measures of ontology complexity

As a test-case, we focus on the Gene Ontology (GO). This ontology is one of the most widely used and actively maintained in the biomedical domain [18]. Among the keys of its success are its continuous evolution and its active curation [19]. Recent efforts focused on improving the modeling of apoptosis and cardiac conduction, and on increasingly using the Web Ontology Language OWL in the GO infrastructure, which in turn supports TermGenie (http://go.termgenie.org/) to automatically place terms in the hierarchy [20].

We investigated whether GO structural complexity increased monotonously over the last five years, as did its size. We focused on the study of nodes' connectivity and of the graph's hierarchy, based mostly on the subsumption relation. In the discussion, we compare our approach to other works focusing on GO evolution.

Resources and Methods

Structure of the gene ontology

The Gene Ontology is a collaborative effort to deliver a species-independent uniform vocabulary for describing gene products [18]. Its classes, also called "GO terms" are organized in three separate branches describing gene products' molecular functions (MF), the biological processes (BP) they participate in and their location in cellular components (CC).

GO also recognises that these classes can have different granularities, i.e. different levels of precision, or be connected by several relations. It organizes them as a directed acyclic graph that supports reasoning (http://www.geneontology.org/GO.ontology.relations.shtml).

Within each branch, the classes are connected by three kinds of relations. The classes are organized in a taxonomy with occasional multiple inheritance along the is a relation which connects a subclass to its superclass (for example, "Carbohydrate metabolic process" (GO:0005975) is a subclass of both "Organic substance metabolic process" (GO:0071704) and "Primary metabolic process" (GO:0044238)). The part of relation connects a part to a whole (for example, "Golgi cisterna" (GO:0031985) is a part of "Golgi stack" (GO:0005795)). The regulates relation connects a regulator process to a regulated process (for example, "Regulation of meiosis" (GO:0040020) regulates "Meiosis" (GO:0007126)). Contrary to the is a and part of relations, regulates has two more specific subrelations: positively regulates and negatively regulates.

Table 1. Ontology complexity metrics.

Ontology aspect	Metrics	Scope	Definition
Size	$\|C_{GO}\|$	Global	Number of classes in GO
	$\|R_{isa}\|$	Global	Number of is a relations in GO
	$\|R_{partof}\|$	Global	Number of part of relations in GO
	$\|R_{regulates}\|$	Global	Number of regulates relations in GO
Connectivity	Average degree	Local	$2*(\|R_{isa}\|+\|R_{partof}\|+\|R_{regulates}\|)/(\|C_{GO}\|)$
	Av. nb is a	Local	$(\|R_{isa}\|)/(\|C_{GO}\|)$
	Av. nb part of	Local	$(\|R_{partof}\|)/(\|C_{GO}\|)$
	Av. nb regulates	Local	$(\|R_{regulates}\|)/(\|C_{GO}\|)$
Hierarchy	Proportion of leaves	Global	(Nb of classes with no subclasses)/$(\|C_{GO}\|)$
	Av. height	Local	$\sum$(max length of path from node to leaf)/$(\|C_{GO}\|)$
	Av. depth	Local	$\sum$(max length of path from node to root)/$(\|C_{GO}\|)$

Description of the metrics used to quantify the complexity variations of an ontology. The definitions are given for GO and can be adapted to BP, CC and MF.
doi:10.1371/journal.pone.0075993.t001

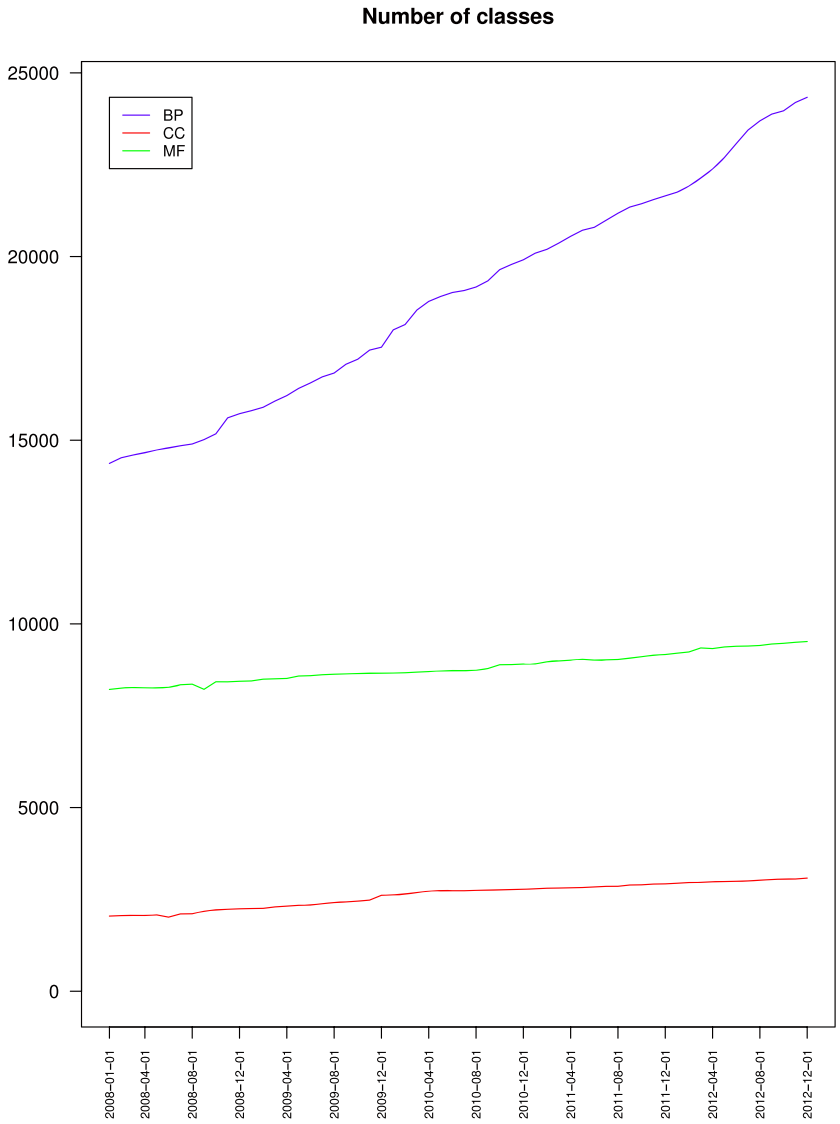


Figure 1. Evolution of the number of classes of the three branches of the Gene Ontology. Biological process (BP), Cellular component (CC) and Molecular function (MF).
doi:10.1371/journal.pone.0075993.g001

This leads to a systematic modeling pattern where each regulation process has two subclasses representing the positive and negative regulation processes (the subclasses of "Regulation of meiosis" (GO:0040020) are "Positive regulation of meiosis" (GO:0045836) and "Negative regulation of meiosis" (GO:0045835)), and each of them is connected to the process they regulate (here, "Meiosis" (GO:0007126)) by either regulates, positively regulates or negatively regulates.

Successive gene ontology versions

We retrieved the 60 successive Gene Ontology monthly releases between January 2008 and December 2012 in the OBO format from the Gene Ontology archives (files gene_ontology_edit.obo.2008-01-01.gz to gene_ontology_edit.obo.2012-12-01.gz at http://www.geneontology.org/ontology-archive/).

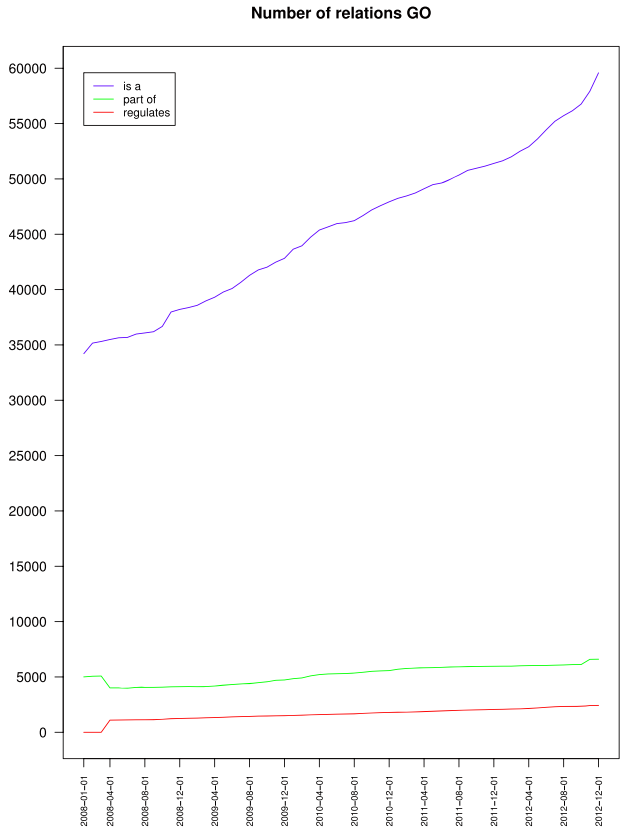
Each of them was converted to the OWL format using Protégé (http://protege.stanford.edu/).

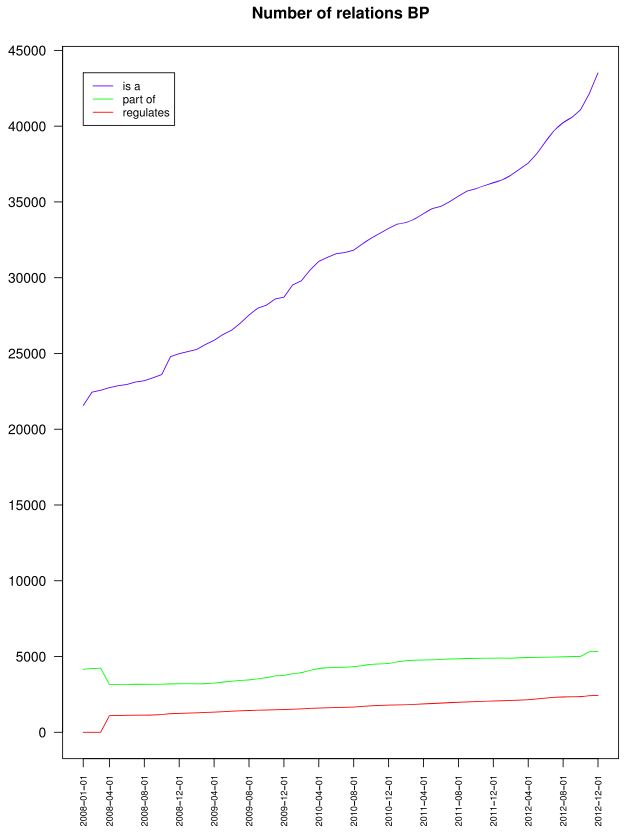
The January and February releases from 2009 appeared to be identical. A personal communication with the Gene Ontology support team confirmed the error and pointed to revision 5.930 from January 31, 2009 from the CVS repository (http://cvsweb.geneontology.org/cgi-bin/cvsweb.cgi/go/).

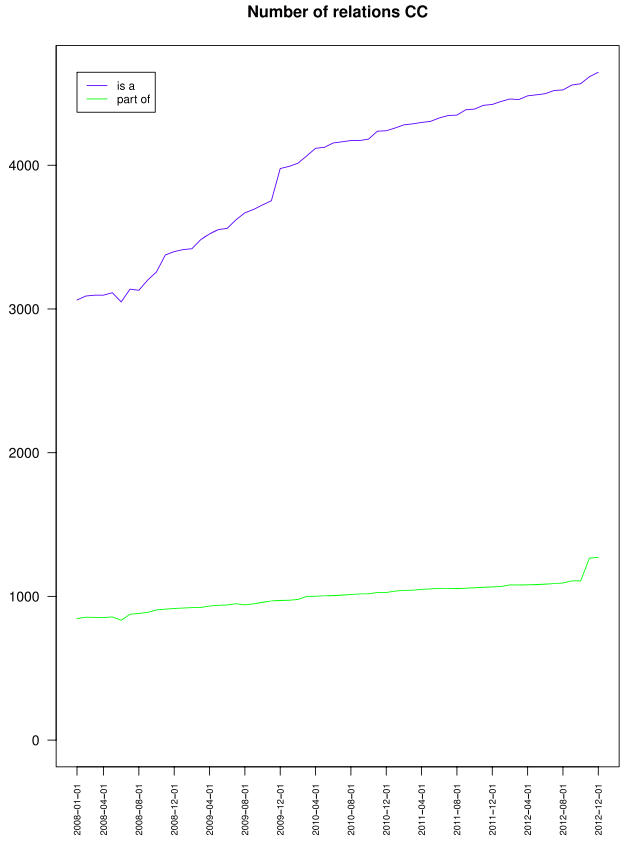
The January 2012 monthly release was not generated. We replaced it by the daily release, which had not changed between 24th December 2011 to 3rd January 2012.

Methods

In order to characterize the evolution of the GO complexity from January 2008 to December 2012, we followed a four-step approach. First, we studied the evolution of the number of classes and relations as a baseline. This gave global indications on the size of the graph. Second, we used several directed acyclic graph (DAG) metrics reflecting the nodes connectivity. This gave local indications on the nodes. Third, we used tree and directed graph hierarchy-related metrics reflecting the graph topological structure. This gave global


Number of relations GO
is a
part of
regulates
60000
55000
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
2008-01-01
2008-04-01
2008-08-01
2008-12-01
2009-04-01
2009-08-01
2009-12-01
2010-04-01
2010-08-01
2010-12-01
2011-04-01
2011-08-01
2011-12-01
2012-04-01
2012-08-01
2012-12-01


Number of relations BP
is a
part of
regulates
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
2008-01-01
2008-04-01
2008-08-01
2008-12-01
2009-04-01
2009-08-01
2009-12-01
2010-04-01
2010-08-01
2010-12-01
2011-04-01
2011-08-01
2011-12-01
2012-04-01
2012-08-01
2012-12-01


Number of relations CC
is a
part of
4000
3000
2000
1000
0
2008-01-01
2008-04-01
2008-08-01
2008-12-01
2009-04-01
2009-08-01
2009-12-01
2010-04-01
2010-08-01
2010-12-01
2011-04-01
2011-08-01
2011-12-01
2012-04-01
2012-08-01
2012-12-01

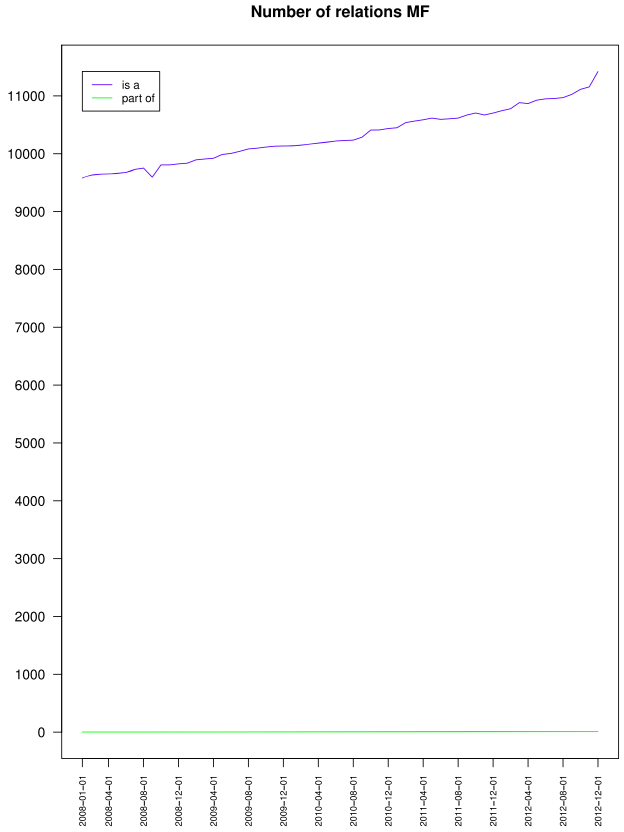

Number of relations MF
is a
part of
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
2008-01-01
2008-04-01
2008-08-01
2008-12-01
2009-04-01
2009-08-01
2009-12-01
2010-04-01
2010-08-01
2010-12-01
2011-04-01
2011-08-01
2011-12-01
2012-04-01
2012-08-01
2012-12-01

Figure 2. Evolution of the number of relations of the Gene Ontology (top left) and its Biological process (top right), Cellular component (bottom left) and Molecular function (bottom right) branches.
doi:10.1371/journal.pone.0075993.g002

indications on the ontology semantics. Fourth, we controlled whether our metrics are able to tell the difference between the real modifications as observed between two successive versions of the ontology, and some random modifications. The idea is that failing to do so would question the relevance of the metrics. We used the February 2010 version of the GO as a baseline. We compared randomly-generated ontology modifications with the March 2010 version in order to study whether or not the previous metrics could discriminate randomly-generated ontology modifications from genuine ones.

During our study, we considered the Gene Ontology both as a whole and by distinguishing its three branches: BP, CC and MF. The branches had different relative sizes. In December 2012, BP represented approximately 66% of the total number of classes, CC represents 8% and MF 26%. The rationale was to detect if some variations of one branch were compensated by some other branch, and to determine if the evolution of the Gene Ontology was uniform among BP, CC and MF.

The modeling pattern for representing process regulation results in each positive or negative regulation relation being systematically subsumed by a regulates relation at the superclass level. In order to avoid counting relations multiple times, we only considered the regulates relation.

Complexity metrics

In this section, we define the graph metrics used to study the evolution of size, connectivity and of topology of GO. Throughout the paper, we used "metrics" to refer to a formula, and "measure" to refer to the value of a metrics. Table 1 summarizes the formal definitions for the metrics used in the first three steps. We adapted the generic framework proposed by Hartung et al. to study the structural changes occurring within ontologies [2]. An ontology modeled as a directed graph is represented by a pair $\langle C,R\rangle$ where C is the set of the classes of the ontology (the nodes of the graph), and R is the set of the typed relations between the classes (the edges). R_{is_a} is the set of the is a relations between classes ($R_{is\_a} \subset R$). Similarly, R_{part_of} and $R_{regulates}$ represent the sets of part of and regulates between the classes ($R = R_{is\_a} \cup R_{part\_of} \cup R_{regulates}$).

The size of an ontology depends on its number of classes and its number of relations. $|C_{GO}|$ represents the number of classes in GO. Likewise, $|C_{BP}|$, $|C_{CC}|$ and $|C_{MF}|$ represent the respective numbers of classes of the BP, CC and MF branches. $|R_{is\_a}|$ represents the number of is a relations in GO. Similarly, $|R_{part\_of}|$ and $|R_{regulates}|$ represent the respective numbers of part of and regulates relations. Branch-specific variations such as $|R_{is\_a,BP}|$ representing the number of is a relations in BP are defined similarly.

Connectivity measures differentiate a sparse graph from a complete graph. The degree of a node is the number of nodes it is directly connected to. Comparing the successive values of the average degree indicated if the graph became more sparse of more dense regardless of the evolution of its size. We used degree-related metrics such as the average number of is a, part of and regulates relations to examine these relations contributions to the average degree.

Ontologies are not only directed acyclic graphs. They also follow a principled hierarchical organization based on the is a relation. Throughout the paper, we used "graph topology" when refering to relations in general, and "graph hierarchy" when refering to metrics taking the semantics into account. In the evolution of an ontology we expect classes to be added at each levels of the hierarchy: close to the root, in the middle, and as leaves (i.e. classes that have no subclasses). Because of inheritance, modifications of an is a relation between two nodes has remote consequences on their descendants and ancestors. To reflect this principled organization of an ontology, we used several hierarchy-related metrics. We computed the proportion of leaves, and nodes'

Table 2. Gene Ontology complexity variations.

	BP			CC			MF		
	Jan 2008	**Dec 2012**	**%**	**Jan 2008**	**Dec 2012**	**%**	**Jan 2008**	**Dec 2012**	**%**
Nb. classes	14,369	24,335	+69.36%	2,046	3,080	+50.54%	8,216	9,520	+15.87%
Nb. relations	25,719	55,341	+115.18%	3,908	5,919	+51.46%	9,583	11,430	+19.27%
Nb. is a	21,563	43,524	+101.85%	3,062	4,647	+51.76%	9,581	11,421	+19.20%
Nb. part of	4,156	5,323	+28.08%	846	1,272	+50.35%	2	9	+350.00%
Nb. regulates	0	2,429		0	0		0	0	
Av. degree	3.58	4.55	+27.05%	3.82	3.84	+0.61%	2.33	2.4	+2.94%
Av. is a	1.5	1.79	+19.18%	1.5	1.51	+0.81%	1.17	1.2	+2.88%
Av. part of	0.29	0.22	−24.37%	0.41	0.41	−0.12%	$2.43E^{-4}$	$9.45E^{-4}$	+288.36%
Av. regulates	0	0.1		0	0		0	0	
Prop. leaves	0.55	0.53	−3.24%	0.76	0.78	+2.71%	0.8	0.8	−0.35%
Av. depth	6.22	7.29	+17.16%	4.97	4.79	−3.46%	5.50	5.62	+2.20%
Max. depth	13	15	+23.08%	10	10	0%	14	15	+7.14%
Av. height	0.89	0.97	+9.19%	0.45	0.40	−11.86%	0.36	0.37	+3.36%

Proportional variations of ontology metrics for Biological process (BP), Cellular components (CC) and Molecular functions (MF) between January 2008 (reference) and December 2012.
doi:10.1371/journal.pone.0075993.t002

Table 3. Proportions of classes for the three Gene Ontology branches.

	BP		CC		MF	
	Classes	% GO	Classes	% GO	Classes	% GO
Jan 2008	14,369	58.34%	2,046	8.31%	8,216	33.36%
Dec 2012	24,335	65.89%	3,080	8.34%	9,520	25.78%

Proportions of total number of Gene Ontology classes for Biological process (BP), Cellular components (CC) and Molecular functions (MF) between January 2008 and December 2012.
doi:10.1371/journal.pone.0075993.t003

Table 4. Proportions of relations for the three Gene Ontology branches.

	BP		CC		MF	
	Relations	% GO	Relations	% GO	Relations	% GO
Jan 2008	25,719	65.59%	3,908	9.97%	9,583	24.44%
Dec 2012	55,341	76.13%	5,919	8.14%	11,430	15.72%

Proportions of total number of Gene Ontology relations for Biological process (BP), Cellular components (CC) and Molecular functions (MF) between January 2008 and December 2012.
doi:10.1371/journal.pone.0075993.t004

average height and average depth. The height of a node is the maximum length of the paths from a leaf to this node. It represents how far a node is from the leaves. The depth of a node is the maximum length of the paths from this node to a root. It represents how far a node is from the root.

Generation and analysis of the random ontologies

We studied if the previous metrics could discriminate randomly-generated ontology modifications from genuine ones. Based on the February 2010 version of GO, we generated fifty simulated ontologies by adding randomly the same numbers of classes and relations. The proportions were respected for BP, CC and MF (e.g. there were 395 classes added to BP in March 2010, so we randomly added 395 classes to BP in each of the fifty simulated ontologies). For each simulation and for BP, CC and MF separately, we created the classes to be added and randomly selected a parent for each of them (thus generating as many random is a relations as classes to be added). We then created the remaining random is a relations, and the random part of and regulates relations. Note that a random class can be created as a subclass of another previous random class, forming a new branch of the hierarchy.

We compared the simulated values with the value observed in March 2010 for average depth, average height and proportion of leaves. The null hypothesis was "There is no statistically significant difference between the measured values of the randomly-generated ontologies and the value observed between the February and March 2010 version of GO". We performed two-sided Student's t-tests with an α parameter of 0.05 using R version 3.0.0.

Results

Spreadsheets containing the results are available as supplementary files.

S1-geneOntology-complexityEvolution-monthly.ods contains the analysis of the sixty Gene Ontology monthly releases between January 2008 and December 2012.

S2-geneOntology-enrichmentSimulations.ods contains the analysis of the fifty simulated random ontologies.

Variations of number of classes and relations

Figure 1 and Figure 2 show that the number of classes and of relations increased monotonously but at different rates during the time of study.

Table 2 shows that the number of classes increased by 50% for GO, 69% for BP, 51% for CC and 16% for MF between January 2008 and December 2012. These different growth rates modified the relative importance of the three branches. Over the study period, Table 3 shows that the proportion of BP classes increased from 58% to 66% of the Gene Ontology, stayed around 8% for CC and decreased from 33% to 26% for MF. Meanwhile, the number of relations increased by 85% for GO, 115% for BP, 51% for CC and 16% for MF. Table 4 shows that the proportion of BP relations increased from 66% to 76% of the Gene Ontology and decreased from 10% to 8% for CC and from 24% to 16% for MF.

At this point, our results confirm the initial impression by OnEX that the Gene Ontology complexity increased monotonously as a whole as well as for its three branches, and that BP was the branch with the fastest growth, which explained why CC and MF were proportionally decreasing.

Variations of connectivity

The number of relations increased, but so did the number of classes. We investigated whether the number of relations increased proportionally more (the graph became denser) or less (the graph became more sparse) than the number of classes. The previous results indicate that between January 2008 and December 2012, the number of relations increased proportionally more than the number of classes for BP, whereas both number increased by similar proportions for CC and MF. We wanted to know if this trend was regular and uniform for the three relations is a, part of and regulates.

Figure 3 presents the evolution of the average degree of a node for BP, CC and MF. It shows that the average degree of a node was around 4 for BP and CC, and around 2.3 for MF.

Figure 3 also shows that over time, the average degree of a node increased monotonously for BP, decreased slightly for CC with some local variations and a sharp increase in November 2012, and remained stable for MF, which completes the previous observations.

Figure 4 and Table 2 present the contributions of the is a, part of and regulates relations to a node's average degree. It shows that the average number is a associated to a node increased for BP but remained stable for CC and MF. The average number of part of associated to a node decreased for BP, was stable for CC and increased slightly for MF. The average number of regulates associated to a node increased for BP.

Overall, these results indicate (1) that GO branches had different connectivity and different variations of connectivity, and (2) that inside a branch the various relations also had different variations.

Variations of hierarchy

Figure 5 presents the variations of the proportion of leaves for GO and its three branches. It shows that the proportion of leaves decreased for BP from 55% to 53.1%, increased for CC from 75.5% to 77.7% and remained stable for MF around 80%. The three branches had different proportions of leaves and different

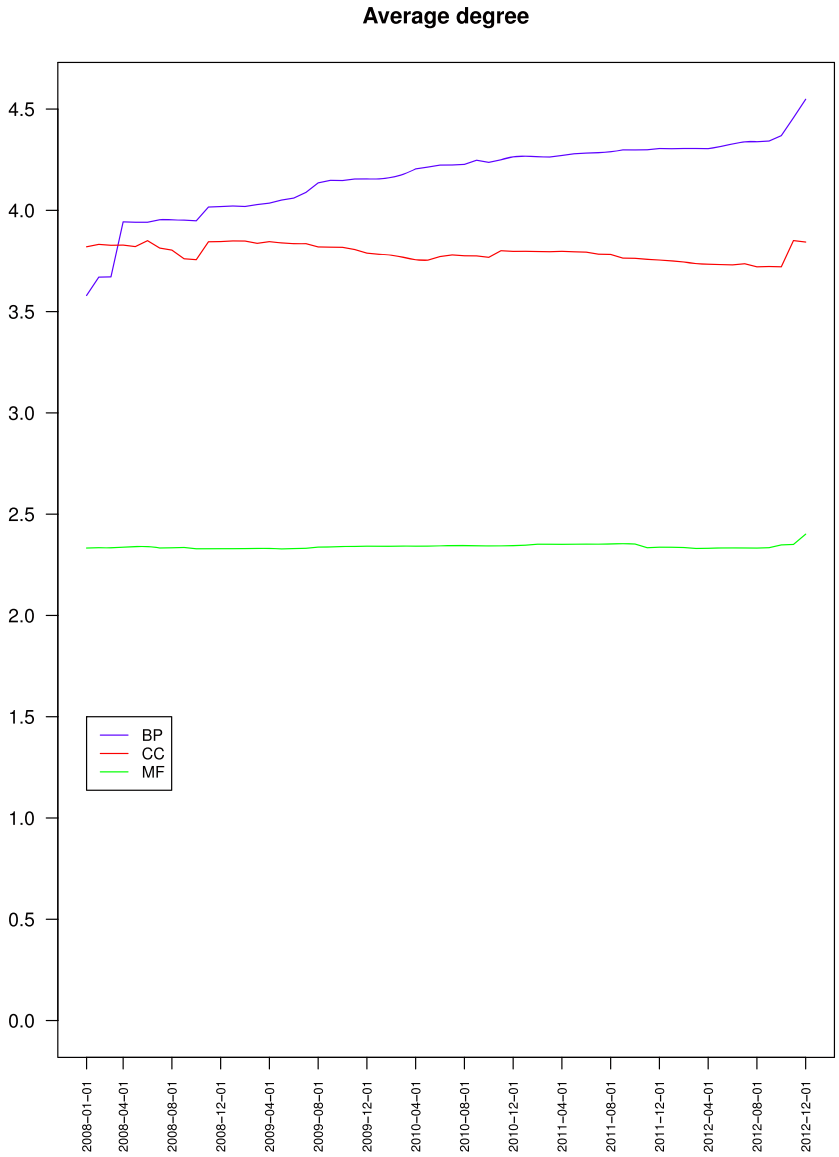


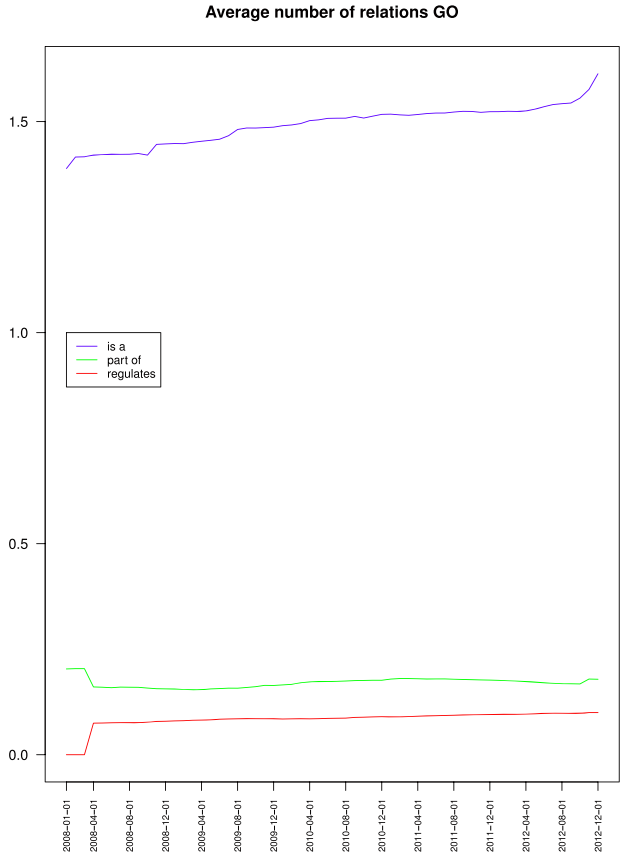
Figure 3. Evolution of the average degree of the nodes of the three branches of the Gene Ontology. Biological process (BP), Cellular component (CC) and Molecular function (MF).
doi:10.1371/journal.pone.0075993.g003

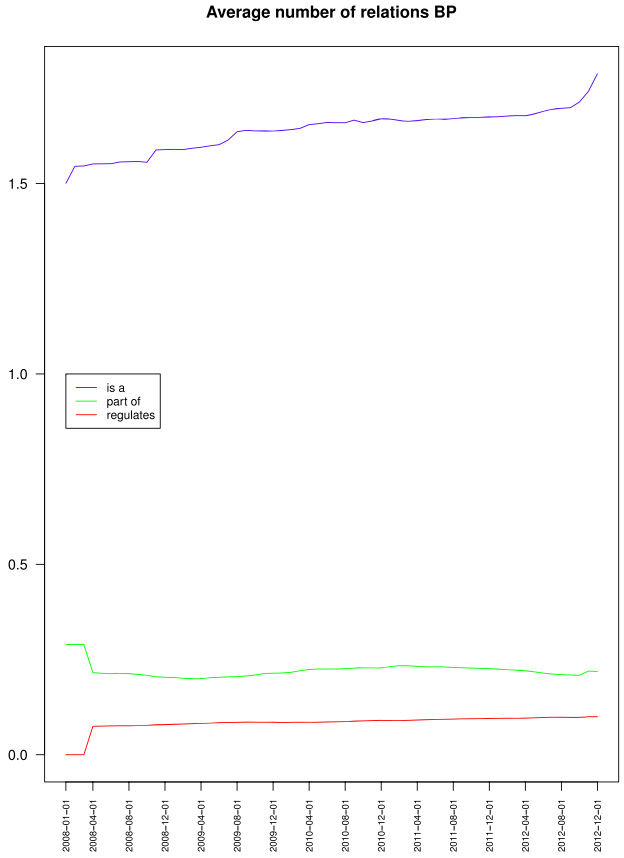
variation patterns. This suggests that the new classes added to BP mostly belong to the intermediate levels of the taxonomy, whereas those added to CC and MF were mostly leaves (maintaining a proportion of 70% to 80% of leaves as the number of classes increases requires that 70% to 80% of the new classes are also leaves).

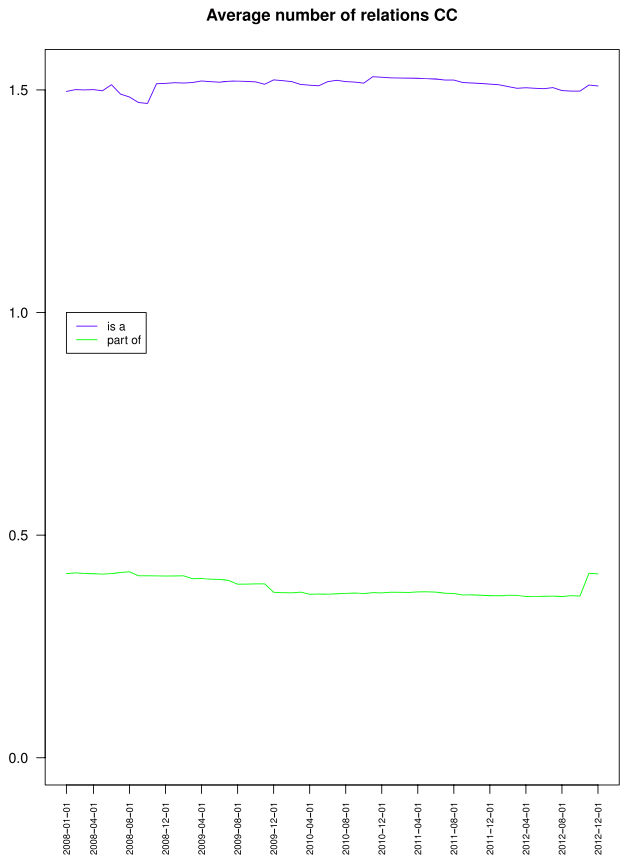
Figure 6 presents the variations of the average height of the nodes from GO and its three branches. It shows that nodes average height increased globally for BP but has been mostly stable since June 2009, decreased for CC and remained mostly stable for MF, which confirms the indications of Figure 5.

Table 2 shows that the maximum depth increased slightly from 13 to 16 for BP, remained at 10 for CC and increased from 14 to 15 for MF. Figure 7 presents the variations of the average depth of the nodes from GO and its three branches. It shows that nodes average depth increased for BP, and remained mostly stable for MF, which confirms the observations of Figures 5 and 6. The fact that for BP both the average depth and the average height increased reinforces the idea that most of the new BP classes were not leaves (or the average height would have decreased), but were parents or ancestors of leaves (because the average distance to a leaf was 0.97) at least 7 edges away from the root (because the average distance to the root increased from 6.2 to 7.3). Figure 7 also shows that the average depth remained mostly stable for CC until March 2012, when it dropped. Together with Figures 2, 5 and 6, this indicates that the new classes added to CC were mostly leaves, and were siblings of existing leaves so that depth was not affected. The March 2012 drop cannot be explained by the variations of number of classes nor of relations or leaves. This suggests some reorganization of the classes hierarchy.

Figures 5, 6 and 7 also compare the relative values of BP, CC and MF proportion of leaves, average height and average depth.


Average number of relations GO
1.5
1.0
0.5
0.0
is a
part of
regulates
2008-01-01
2008-04-01
2008-08-01
2008-12-01
2009-04-01
2009-08-01
2009-12-01
2010-04-01
2010-08-01
2010-12-01
2011-04-01
2011-08-01
2011-12-01
2012-04-01
2012-08-01
2012-12-01


Average number of relations BP
1.5
1.0
0.5
0.0
is a
part of
regulates
2008-01-01
2008-04-01
2008-08-01
2008-12-01
2009-04-01
2009-08-01
2009-12-01
2010-04-01
2010-08-01
2010-12-01
2011-04-01
2011-08-01
2011-12-01
2012-04-01
2012-08-01
2012-12-01


Average number of relations CC
1.5
1.0
0.5
0.0
is a
part of
2008-01-01
2008-04-01
2008-08-01
2008-12-01
2009-04-01
2009-08-01
2009-12-01
2010-04-01
2010-08-01
2010-12-01
2011-04-01
2011-08-01
2011-12-01
2012-04-01
2012-08-01
2012-12-01

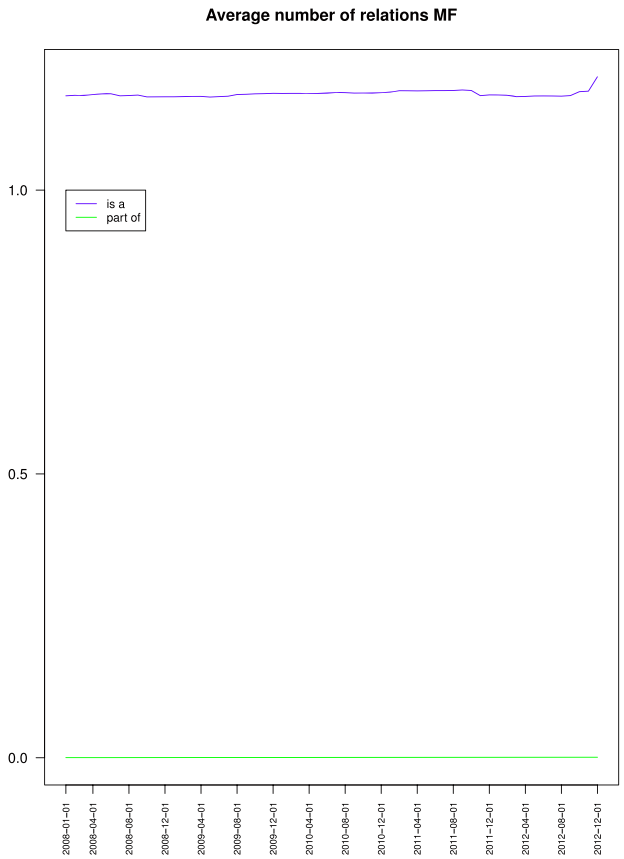

Average number of relations MF
1.0
0.5
0.0
is a
part of
2008-01-01
2008-04-01
2008-08-01
2008-12-01
2009-04-01
2009-08-01
2009-12-01
2010-04-01
2010-08-01
2010-12-01
2011-04-01
2011-08-01
2011-12-01
2012-04-01
2012-08-01
2012-12-01

Figure 4. Contributions of the is a, part of and regulates relations to a node's average degree for the Gene Ontology (top left) and its three branches Biological process (top right), Cellular component (bottom left) and Molecular function (bottom right).
doi:10.1371/journal.pone.0075993.g004

The three metrics reflecting the semantics of the ontology exhibited a similar pattern with CC and MF having similar values compared to BP. This should be contrasted with connectivity metrics from Figure 4 where BP and CC had similar average degree values, compared to MF. Interestingly, CC was similar to BP from a connectivity point of view, and similar to MF from a semantic structure point of view. The similar connectivity of BP and CC is reinforced by the fact that both rely on is a and part of relations, whereas MF almost exclusively uses is a (Table 2).

Comparison with random ontology enrichment

The previous results about the local variations of node connectivity and the global variations of the graph structure showed some fairly monotonous trends for BP, CC and MF. We investigated if these trends were the result of the sole increase of classes and relations. We studied if the previous metrics could discriminate randomly-generated ontology modifications from genuine ones. Table 5 presents the variation of the number of classes and relations between the February and March 2010 versions of the GO, and the average of these metrics on the fifty simulated ontologies.

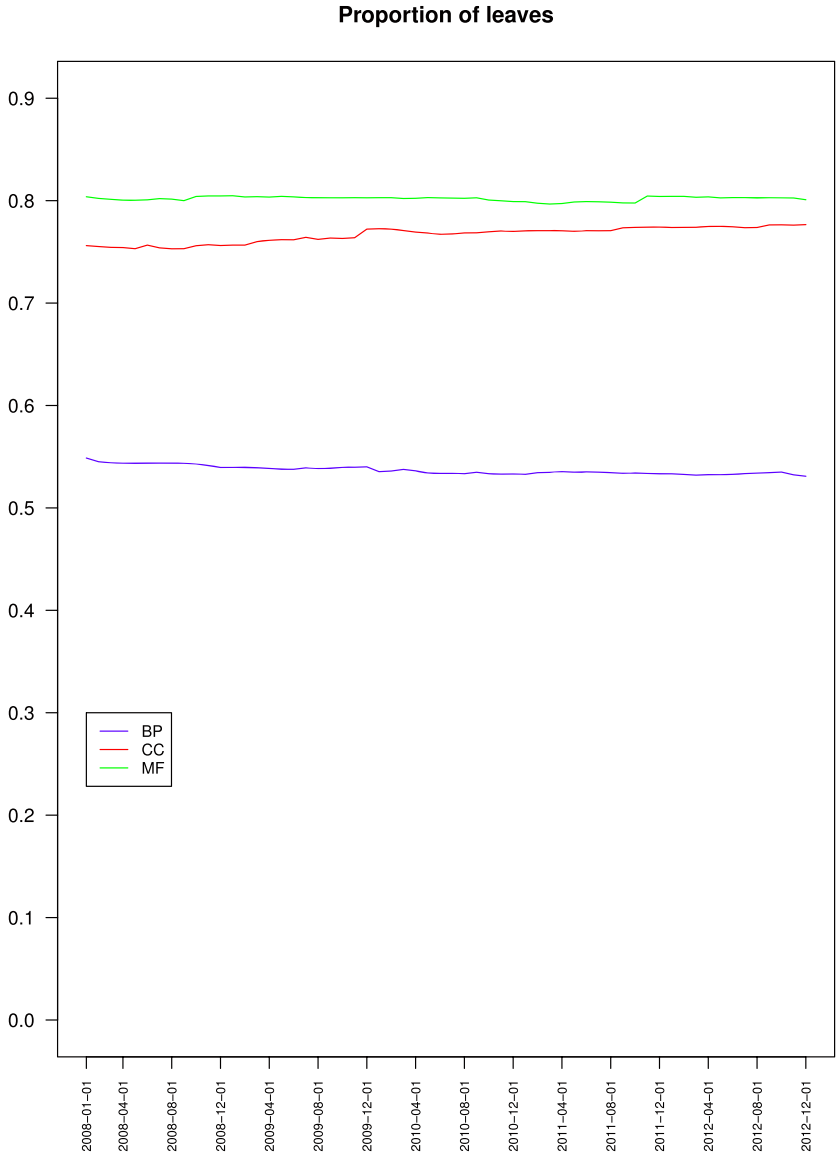


Figure 5. Variations of the proportion of leaves for the Gene Ontology three branches. Biological process (BP), Cellular component (CC) and Molecular function (MF).
doi:10.1371/journal.pone.0075993.g005

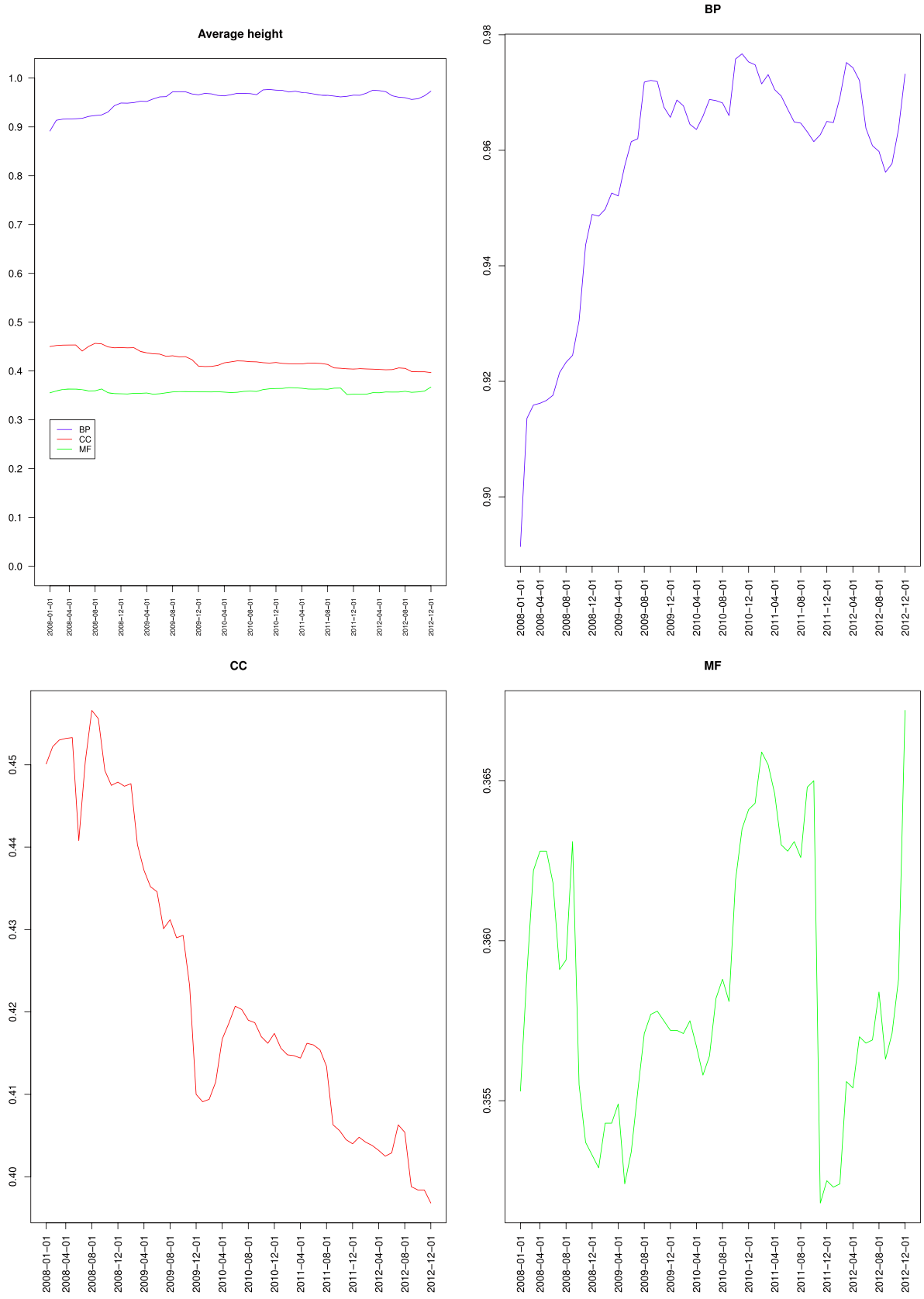


Figure 6. Variations of the average height of the nodes from the Gene Ontology: together (top left), Biological process (top right), Cellular component (bottom left) and Molecular function (bottom right).
doi:10.1371/journal.pone.0075993.g006

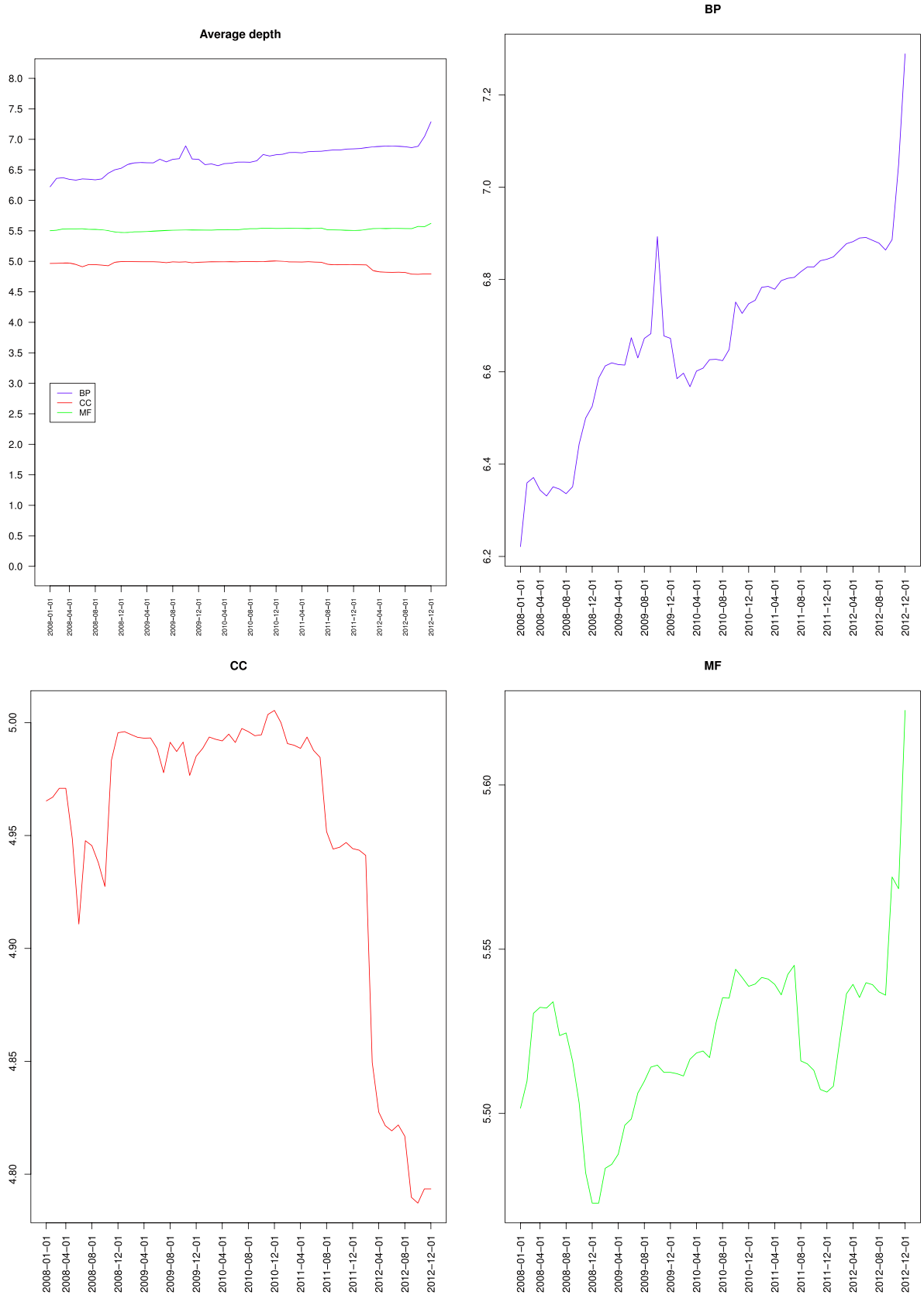


Figure 7. Variations of the average depth of the nodes from the Gene Ontology: together (top left), Biological process (top right), Cellular component (bottom left) and Molecular function (bottom right).
doi:10.1371/journal.pone.0075993.g007

Table 5. Simulated evolution of the three Gene Ontology branches between February and March 2010.

	BP			CC			MF		
	Feb. 2010	**Mar. 2010**	**simul.**	**Feb. 2010**	**Mar. 2010**	**simul.**	**Feb. 2010**	**Mar. 2010**	**simul.**
Nb. classes	18,149	18,544	18,544	2,643	2,688	2,688	8,670	8,687	8,687
Nb. is a	29,796	30,507	30,507	4,014	4,065	4,065	1,047	1,067	1,067
Nb. part of	3,928	4,090	4,090	979	1,000	1,000	4	7	7
Nb. regulates	1,542	1,580	1,580	0	0	0	0	0	0
Av. depth	6.597	6.567	7.275	4.994	4.993	5.022	5.511	5.517	5.513
Av. height	0.968	0.965	1.104	0.409	0.411	0.433	0.357	0.358	0.360
Prop. leaves	0.536	0.538	0.525	0.772	0.771	0.761	0.803	0.802	0.801

Variations of ontology metrics for Biological process (BP), Cellular components (CC) and Molecular functions (MF) between February and March 2010, compared to the average of fifty randomly-enriched simulations.
doi:10.1371/journal.pone.0075993.t005

Connectivity metrics are based on the average number of relations. Therefore, they were not affected by the simulations.

Figures 8, 9 and 10 present the proportion of leaves, average height and average depth of the simulations compared to the March 2010 version of GO.

BP simulations had fewer leaves, higher average depths and higher average heights than GO. CC simulations had fewer leaves and higher average heights than GO, but similar average depths. MF simulations had fewer leaves than GO, but similar average heights and average depths.

Table 6 presents the p-values of the Student's t-tests. All the tests showed a statistically significant difference between the simulated and the observed values, except for the average depth in MF. For MF, the fact that the average height increased more in the simulated ontologies than in the March 2010 version of GO, and that the proportion of leaves decreased more in the simulations suggests that the simulated classes were mostly added as non-leaves. The lack of statistically-significant difference of average depth is difficult to interpret, specially because there was a difference of average height. Possible factors are the small number of modifications for MF (but this argument also hold for the other measures), or the structure of MF hierarchy.

Together, the random ontology enrichment results confirm that the average depth, average height and proportion of leaves can discriminate randomly-generated ontology modifications from genuine ones. The differences between BP, CC and MF also confirm the previous observations that the three branches have different hierarchical organizations, and different evolutions. The lower number of leaves observed in BP, CC and MF for the simulations were consistent with the higher average heights: if randomly-added classes are not leaves, they are at least one edge away from the leaves; since each branch average height was lower than 1, these classes tend to increase the average height. The difference between BP depth and height variations on the one hand and CC and MF variations on the other hand can be explained by the structural differences between the former and the last two. BP has a smaller proportion of leaves than CC and MF so that randomly-added classes are less likely to be leaves than for CC or MF. Interestingly, Pesquita et al. also observed that for the GO, the refinement of CC and MF occurs mostly via single insertions, whereas in BP, groups of related classes are inserted together [21].

These simulations also confirm that in complex graph structures like ontologies, a small number of changes in the topology can have dramatic consequences on the overall hierarchy. Applications based on approaches such as term enrichment are highly sensitive to such modifications because the annotations are propagated to the ancestors [22–25].

Discussion

In this section, we first survey related GO-specific works. We then discuss the practical applications of our study. Finally, we discuss how our approach can be generalized to other ontologies and other metrics.

GO-specific approaches

Several studies analyzed the evolution of the GO from different perspectives.

Park et al. developed visualization methods based on a color-coded layered graph to highlight the changes between two versions of GO [26]. Hartung et al. improved the idea with CODEX, that determines a compact diff based on semantic changes [27]. Both approaches focus on change visualization but leave the interpretation of the modifications to the user.

Leonelli et al. characterized the reasons of the changes. They identified five circumstances warranting changes in the GO by curators: (1) the emergence of anomalies within GO; (2) the extension of the scope of GO; (3) the divergence in how terminology is used across user communities; (4) new discoveries that change the meaning of the terms used and their relations to each other; and (5) the extension of the range of relations used to link entities or processes described by GO terms [28]. They focus on improving the way the GO represents biological knowledge but leave the determination of the quality change to the curators and do not measure it.

Köhler et al. proposed a systematic method to analyze the quality of terms definitions [29]. Verspoor et al. developed a transformation-based automatic clustering method for detecting similar terms that use different linguistic conventions [30]. Both approaches focus on the classes names or textual definitions but do not consider the relations among the classes. Mungall et al. proposed an automatic reasoning-based approach using logical definitions for classes and mappings to external ontologies that detects potentially missing and incorrect classes and relationships [31]. It should be noted that even if logical definitions are assigned to all new regulation classes as of January 2010, processing all the previous classes is an ambitious ongoing task. Alterovitz et al. proposed an information theory-based approach to automatically organize the structure of GO and optimize the distribution of the information within it [32]. Faria et al. proposed an association

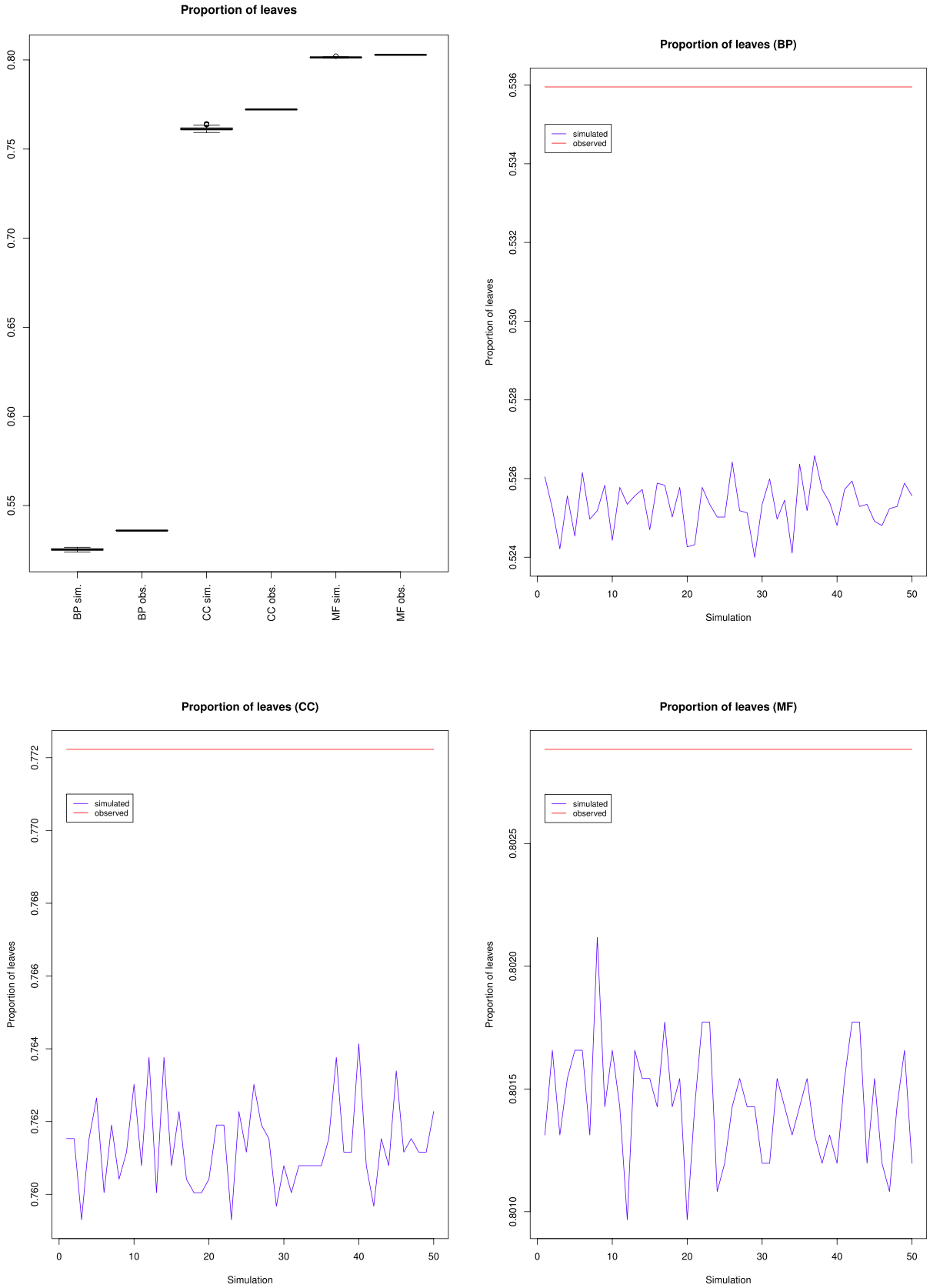


Figure 8. Proportion of leaves for the fifty simulated ontologies, compared to the value for the March 2010 version of the Gene Ontology (red line).
doi:10.1371/journal.pone.0075993.g008

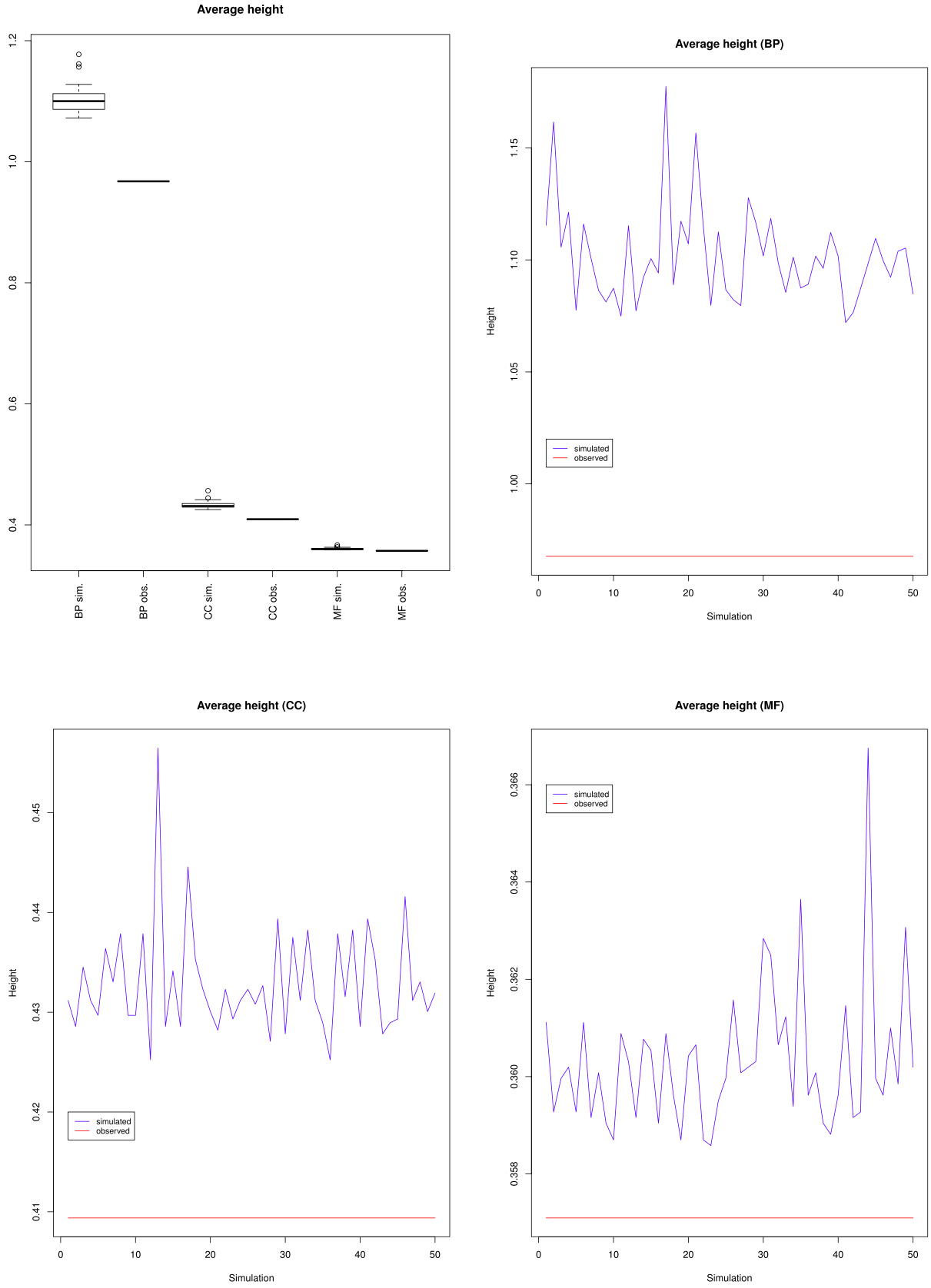


Figure 9. Average classes' heights for the fifty simulated ontologies, compared to the value for the March 2010 version of the Gene Ontology (red line).
doi:10.1371/journal.pone.0075993.g009

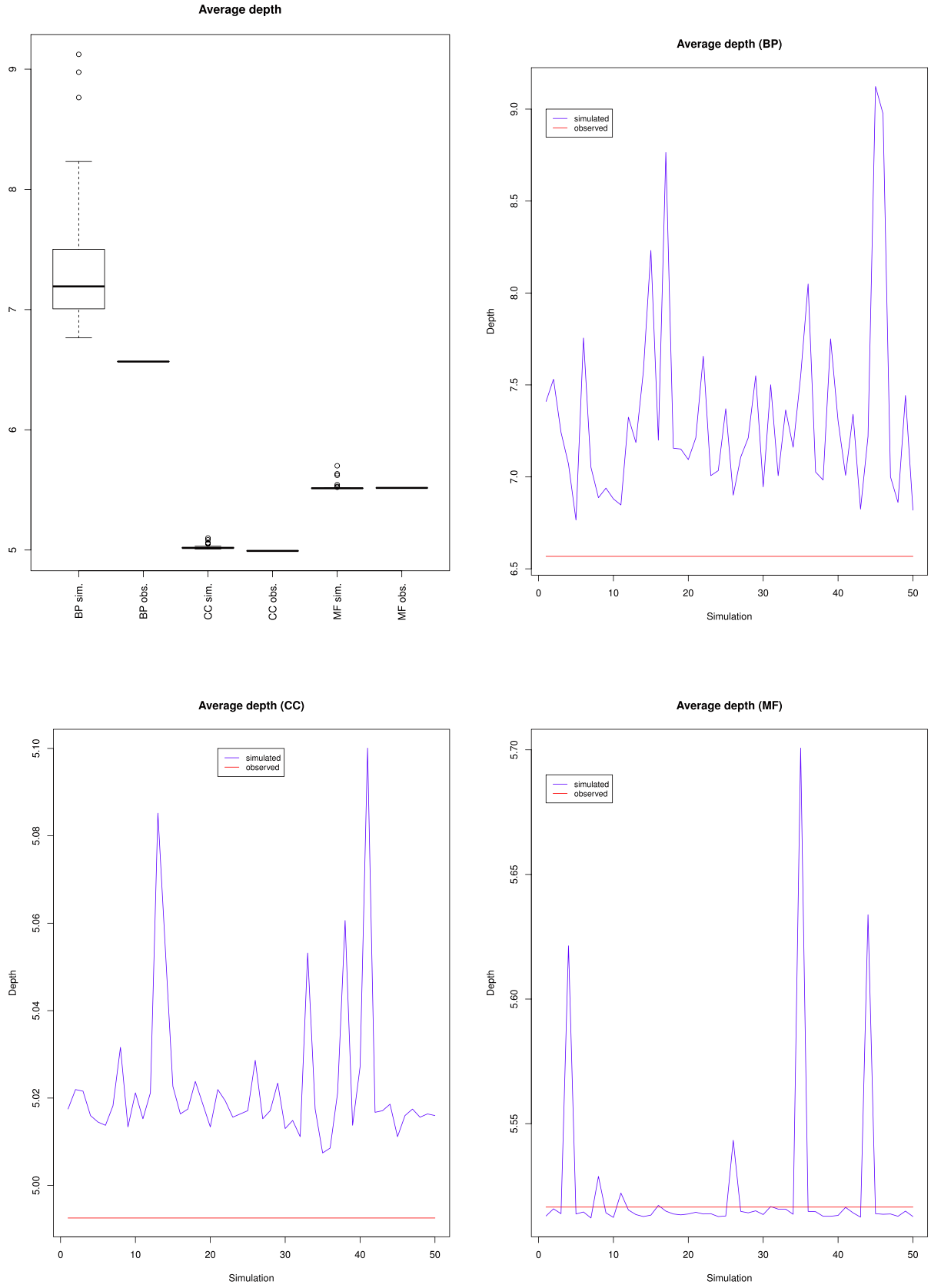


Figure 10. Average classes' depths for the fifty simulated ontologies, compared to the value for the March 2010 version of the Gene Ontology (red line).
doi:10.1371/journal.pone.0075993.g010

Table 6. Comparison of the fifty randomly enriched ontologies with the March 2010 version of Gene Ontology.

	BP	CC	MF
av. depth	$7.657E^{-14}$	$2.4E^{-16}$	0.1643
av. height	$<2.2E^{-16}$	$<2.2E^{-16}$	$<2.2E^{-16}$
proportion leaves	$<2.2E^{-16}$	$<2.2E^{-16}$	$<2.2E^{-16}$

P-value of Student's t-tests comparing the fifty randomly enriched ontologies with the March 2010 version of Gene Ontology.
doi:10.1371/journal.pone.0075993.t006

rule-based algorithm for identifying implicit relationships between molecular function terms [33]. Other works focused on the quality of terms definitions [29] and on the detection of semantic inconsistencies of gene annotations [34]. Gross et al. studied to what extent modifications of the GO and of gene annotations databases impacted the result of term enrichment analyses that describe experimental data by sets of GO terms [35]. They demonstrated that the "changes are unequally distributed and cluster in regions representing specific topics". Interestingly, they also observed that these changes do not necessarily modify the result of term enrichment analyses since the terms are often semantically related. Our results indicated that for BP, most modifications occurred deep into the hierarchy, so it is also possible that term enrichment analyses return sets of more general GO terms that are more stable. Loguercio et al. proposed a task-based approach to examine the completeness and utility of GO annotations for gene enrichment analysis [24]. It should be noted that over time, both gene annotations (i.e. the set of GO terms associated to gene products) and the GO itself evolve simultaneously. They focused on the quality of annotations, whereas we focused on GO proper. Moreover, as stated in the background section, the metrics of complexity we used are intrinsic values that are task-independent.

Ceusters performed an extensive evolutionary terminology auditing [36] of the GO between 2001 and 2007 for measuring to what extent the structure of a terminology mimics reality. This avoids mistakes, some of which are not eliminated by automatic reasoning. He reports that the quality of the BP, CC and MF branches of the GO increased continuously over time, with MF having consistently the highest quality. He also observed a 'high correlation (0.95) between the increase in size of the GO as a whole and the quality scores'. This should be contrasted with our results (admittedly over a different period) showing that the complexity increased for BP, decreased slightly for CC and remained stable for MF.

Pesquita and Couto proposed a semi-automatic approach for change capture, i.e. the identification of the areas of an ontology that need to be changed [21]. They applied it to 6-months spaced snapshots of the GO over the 2005–2010 period to study whether their framework could predict the portions that would be extended. Their focus was on the analysis of the new classes and relations. It relied on (1) the depth of new classes, (2) the number of new classes that are children of (former) leaves, and (3) the number of new classes that are children of existing classes vs. of newly added classes. This allowed to determine the general direction of refinement (i.e. if new classes provide a finer description or cover a new domain) and whether new classes are inserted individually or as parts of a new branch. They observed that in BP, CC and MF, the majority of new subclasses are added as children of non-leaf classes. They also observed that the refinement of CC and MF occurs mostly via single insertions, whereas in BP, groups of related classes are inserted together. Their observations are compatible with our results. It should be noted that their approach focuses on the analysis of the features of the new classes, whereas we studied BP, CC and MF globally and focused on the consequences of the changes (not just the additions) on the ontology itself. Therefore, we believe the two approaches complement each other.

Practical applications

The main consequences of our results concern people maintaining GO annotations, as well as developpers of data analysis methods based on the GO.

The regular addition of leaves or of classes close to leaves for BP and CC indicates that over time, more precise terms were being added to the GO hierarchy. Some of the former annotations that refer to the parents of these new classes could be transferred to the new classes. Because of the rule of annotations propagation to the ancestors, the former annotations would remain valid, but this would result in a gain in annotation precision. With the OnEX web application, Hartung et al. proposed a mechanism capable of semi-automatic migration of outdated annotations [4]. Our results indicate that the addition of new low-level classes (mostly for BP and CC) has potential implications on former annotations, whereas higher level classes (mostly for MF) represent previously undescribed topics. The latter situation is not compatible with the OnEX semi-automatic migration approach. Ideally, experts should decide whether these new high-level annotations are suitable for existing entities such as gene products.

The parallel evolution of the GO and of annotations databases has consequences on the results of data analysis studies [37] as well as on the evaluation of GO-based data analysis methods [38–40]. Gillis et al. reported that "GO annotations are stable over short period of time", but also that "genes can alter their functional identity with 20% of gene not matching to themselves (by semantic similarity) after two years" [25]. The direct implication is that all the results of analyses based on the GO should be re-assessed on a regular basis. By showing that complexity increased for BP and CC with the addition of leaves or of classes close to leaves and that MF complexity remained stable with uniform modifications, our study suggests that the conclusions of the previous analyses could remain valid but may actually be improved, although quantifying this assumption would be a separate work. Similarly, the respective performances of GO-based data analysis methods should be re-evaluated on a regular basis.

These metrics could be integrated into at least three kinds of future applications. First, they could easily be integrated into ontology-development tools such as Protégé or OBOEdit. However, not all users may have the need to monitor such metrics. Furthermore, comparing the measures when only a few changes have been made may make it harder to identify general trends. We also computed the measures on daily snapshots of GO from July 2009 to July 2012 and observed successive increases and decreases on all values. The second option would then be to integrate our metrics on top of the ontology version control system. We have seen that computing the measures between commits is not very informative, whereas comparing their evolution between releases (i.e. when the curators judge that a set of commits achieved a meaningful goal) makes more sense. The third alternative would be to integrate our metrics into ontology repositories such as Onex (http://dbserv2.informatik.uni-leipzig.de:8080/onex/or Bioportal (http://bioportal.bioontology.org/). This solution is user-oriented, whereas the second one was curator-oriented.

Generalization

Our approach relies on classic DAG metrics, none of which is GO-specific. Therefore, our approach is readily applicable to any other ontology. It has the advantage of genericity, but the drawback is that it would probably ignore some ontologies peculiarities (e.g. the positive and negative regulation pattern, which has an impact on the nodes' degree). These would have to be taken into account when interpreting the results.

This argument makes the comparison of the values between ontologies questionable (e.g. to determine thresholds or to provide some qualitative interpretation). We advise to focus on the evolution of measures during an ontology lifecycle.

The next challenge will be to propose new ontology complexity metrics capable of taking into account features of semantically-rich languages such as OWL (http://www.w3.org/TR/owl2-primer/): disjontness between classes, the fact that some relations can be transitive or asymmetric, existential and universal restrictions, etc [41,42]. The connectivity and hierarchy-related metrics that we presented only cover a limited portion of the meaning conveyed in ontologies. They see ontologies mostly as taxonomies, i.e. a directed acyclic graph of is a relations. Most current ontologies are in the taxonomy category anyway, so taking these additional features into account would probably have a limited impact. However, one can anticipate that these features will gradually gain acceptance as they make ontology maintenance easier, and support more advanced reasoning [43,44]. Conversely, providing a quantified measurement of their impact on the ontology structure may also help promoting their adoption.

Conclusion

For the Gene Ontology, the number of classes and relations increased monotonously between January 2008 and December 2012. Considering the three branches of the Gene Ontology (Biological process, Cellular component and Molecular functions) independently gave similar conclusions but revealed different growth rates. Connectivity and hierarchy-related metrics provided additional insights into the ontology complexity. They revealed different patterns in terms of values as well as of evolution.

Graph-related metrics such as the average degree of a node provided additional information about the ontology connectivity. For the Gene Ontology, BP and CC had similar average degrees, superior to that of MF. The analysis of the variations of nodes average degree showed that during the study period, the connectivity of BP nodes increased, while it slightly decreased for CC and remained stable for MF. It also showed that the CC decrease could be attributed to the number of part of relations increasing less than the number of CC classes.

Hierarchy-related metrics such as the proportion of leaves, the average depth and the average height of nodes provided information about the semantics. For the Gene Ontology, CC and MF had similar proportions of leaves, average depths and average heights, that were superior to that of BP for the proportion of leaves, and inferior to BP average depth and average height. The proportion of leaves decreased for BP, increased for CC and remained stable for MF. The nodes average height increased for BP, decreased for CC and remained mostly stable for MF. The nodes average depth increased for BP, remained mostly stable for CC until March 2012 and then decreased, and remained mostly stable for MF. These measures also indicated that most of the classes added to BP were not leaves but were in the lowest part of the hierarchy, whereas most of the classes added to CC were leaves and siblings of existing leaves, and that MF growth was rather uniform. Eventually, hierarchy-related measures could distinguish the actual GO evolution from the random addition and removal of classes and relations.

Overall, for the Gene Ontology, the results showed that the three branches Biological Process, Cellular Component and Molecular Function have to be considered separately when studying the evolution of the Gene Ontology complexity. The number of classes and relations increased monotonously for all branches. Our results show that the changes operated by Gene Ontology curators between monthly releases impact both the ontology size and the ontology complexity. Node connectivity increased monotonously for BP, decreased globally with several local extrema for CC and was stable for MF, with BP and CC having similar profiles compared to MF. Concerning the hierarchy, average depth and average height increased for BP, decreased for CC and was stable for MF, with CC and MF having similar profiles compared to BP. These results indicate that BP was the most dynamic branch which complexity increased, that CC was refined with the addition of leaves providing a finer level of annotations but complexity decreased, and that MF experienced a stable and uniform growth.

Supporting Information

Spreadsheet S1 Analysis of the Gene Ontology monthly releases 2008–2012. (S1-geneOntology-complexityEvolution-monthly.ods) contains the analysis of the sixty Gene Ontology monthly releases between January 2008 and December 2012.
(ODS)

Spreadsheet S2 Analysis of the simulated random ontologies. (S2-geneOntology-enrichmentSimulations.ods) contains the analysis of the fifty simulated random ontologies.
(ODS)

Acknowledgments

Ronald Cornet sparked the inspiration for this work and provided valuable insights.

Author Contributions

Conceived and designed the experiments: OD. Performed the experiments: OD CB NLM. Analyzed the data: OD CB NLM. Wrote the paper: OD CB NLM.

References

1. Bodenreider O, Stevens R (2006) Bio-ontologies: current trends and future directions. Briefings in Bioinformatics 7: 256–274.
2. Hartung M, Toralf K, Rahm E (2008) Analyzing the evolution of life science ontologies and mappings. In: 5th Intl. Workshop on Data Integration in the Life Sciences (DILS) 2008. LNCS 5109.
3. Kirsten T, Gross A, Hartung M, Rahm E (2011) Gomma: A component-based infrastructure for managing and analyzing life science ontologies and their evolution. Journal of biomedical semantics 2: 6.
4. Hartung M, Kirsten T, Gross A, Rahm E (2009) Onex: Exploring changes in life science ontologies. BMC bioinformatics 10: 250.
5. Malone J, Stevens R (2012) Measuring the level of activity in community built bio-ontologies. Journal of biomedical informatics 46: 5–14.
6. Guarino N, Welty C (2002) Evaluating ontological decisions with Ontoclean. In: Communications of the ACM. volume 45, pp. 61–65.
7. Köhler S, Bauer S, Mungall CJ, Carletti G, Smith CL, et al. (2011) Improving ontologies by automatic reasoning and evaluation of logical definitions. BMC bioinformatics 12: 418.
8. Shchekotykhin K, Friedrich G, Fleiss P, Rodler P (2012) Interactive ontology debugging: Two query strategies for efficient fault localization. Web semantics (Online) 12–13: 88–103.

9. Yao L, Divoli A, Mayzus I, Evans JA, Rzhetsky A (2011) Benchmarking ontologies: bigger or better? PLoS computational biology 7: e1001055.
10. Hoehndorf R, Dumontier M, Gkoutos GV (2012) Evaluation of research in biomedical ontologies. Briefings in bioinformatics .
11. Cimino JJ (1998) Auditing the unified medical language system with semantic methods. Journal of the American Medical Informatics Association 5: 41–51.
12. Baorto D, Li L, Cimino JJ (2009) Practical experience with the maintenance and auditing of a large medical ontology. Journal of biomedical informatics 42: 494–503.
13. Cimino JJ, Hayamizu TF, Bodenreider O, Davis B, Stafford GA, et al. (2009) The caBIG terminology review process. Journal of biomedical informatics 42: 571–580.
14. de Coronado S, Wright LW, Fragoso G, Haber MW, Hahn-Dantona EA, et al. (2009) The nci thesaurus quality assurance life cycle. Journal of biomedical informatics 42: 530–539.
15. Gu HH, Wei D, Mejino JLV, Elhanan G (2009) Relationship auditing of the fma ontology. Journal of biomedical informatics 42: 550–557.
16. Meehan TF, Masci AM, Abdulla A, Cowell LG, Blake JA, et al. (2011) Logical development of the cell ontology. BMC bioinformatics 12: 6.
17. Rector AL, Brandt S, Schneider T (2011) Getting the foot out of the pelvis: modeling problems affecting use of snomed ct hierarchies in practical applications. Journal of the American Medical Informatics Association : JAMIA 18: 432–440.
18. The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. Nature genetics 25: 25–29.
19. Bada M, Stevens R, Goble C, Gil Y, Ashburner M, et al. (2004) A short study on the success of the gene ontology. Journal of Web Semantics 1: 235–240.
20. The Gene Ontology Consortium (2012) Gene ontology annotations and resources. Nucleic acids research 41: D530–D535.
21. Pesquita C, Couto FM (2011) Where GO is going and what it means for ontology extension. In: International Conference on Biomedical Ontology ICBO.
22. Khatri P, Draghici S (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics (Oxford, England) 21: 3587–3595.
23. Jin B, Lu X (2010) Identifying informative subsets of the gene ontology with information bottleneck methods. Bioinformatics (Oxford, England) 26: 2445–2451.
24. Loguercio S, Clarke EL, Good BM, Su AI (2012) A task-based approach for large-scale evaluation of the gene ontology. In: IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology, HISB 2012. p. 144.
25. Gillis J, Pavlidis P (2013) Assessing identity, redundancy and confounds in gene ontology annotations over time. Bioinformatics (Oxford, England) .
26. Park JC, Kim Te, Park J (2008) Monitoring the evolutionary aspect of the gene ontology to enhance predictability and usability. BMC bioinformatics 9 Suppl 3: S7.
27. Hartung M, Gross A, Rahm E (2012) CODEX: exploration of semantic changes between ontology versions. Bioinformatics (Oxford, England) 28: 895–896.
28. Leonelli S, Diehl AD, Christie KR, Harris MA, Lomax J (2011) How the gene ontology evolves. BMC bioinformatics 12: 325.
29. Köhler J, Munn K, Rüegg A, Skusa A, Smith B (2006) Quality control for terms and definitions in ontologies and taxonomies. BMC Bioinformatics 7: 212.
30. Verspoor K, Dvorkin D, Cohen KB, Hunter L (2009) Ontology quality assurance through analysis of term transformations. Bioinformatics (Oxford, England) 25: i77–i84.
31. Mungall CJ, Bada M, Berardini TZ, Deegan J, Ireland A, et al. (2010) Cross-product extensions of the gene ontology. Journal of biomedical informatics 44: 80–86.
32. Alterovitz G, Xiang M, Hill DP, Lomax J, Liu J, et al. (2010) Ontology engineering. Nature Biotechnology 28: 128–130.
33. Faria D, Schlicker A, Pesquita C, Bastos H, Ferreira AEN, et al. (2012) Mining GO annotations for improving annotation consistency. PloS one 7: e40519.
34. Park YR, Kim J, Lee HW, Yoon YJ, Kim JH (2011) GOChase-II: correcting semantic inconsistencies from gene ontology-based annotations for gene products. BMC bioinformatics 12 Suppl 1: S40.
35. Gross A, Hartung M, Prüfer K, Kelso J, Rahm E (2012) Impact of ontology evolution on functional analyses. Bioinformatics (Oxford, England) 28: 2671–2677.
36. Ceusters W (2008) Applying evolutionary terminology auditing to the gene ontology. Journal of biomedical informatics 42: 518–529.
37. Yang H, Nepusz T, Paccanaro A (2012) Improving go semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. Bioinformatics (Oxford, England) 28: 1383–1389.
38. Pesquita C, Faria D, Bastos H, Ferreira AE, Falcão AO, et al. (2008) Metrics for go based protein semantic similarity: a systematic evaluation. BMC Bioinformatics 9: S4.
39. Pesquita C, Faria D, Falco AO, Lord P, Couto FM (2009) Semantic similarity in biomedical ontologies. PLoS computational biology 5: e1000443.
40. Wu X, Pang E, Lin K, Pei ZM (2013) Improving the measurement of semantic similarity between gene ontology terms and gene products: Insights from an edge- and ic-based hybrid method. PloS one 8: e66745.
41. Aranguren ME, Bechhofer S, Lord P, Sattler U, Stevens R (2007) Understanding and using the meaning of statements in a bio-ontology: recasting the gene ontology in OWL. BMC bioinformatics 8: 57.
42. Stevens R, Egaña Aranguren M, Wolstencroft K, Sattler U, Drummond N, et al. (2007) Using OWL to model biological knowledge. International Journal of Human Computer Studies 65: 583–594.
43. Golbreich C, Horridge M, Horrocks I, Motik B, Shearer R (2007) OBO and OWL: Leveraging semantic web technologies for the life sciences. In: Proceedings of the 6th International Semantic Web Conference (ISWC 2007). volume 4825 of Lecture Notes in Computer Science, pp. 169–182.
44. Jupp S, Stevens R, Hoehndorf R (2012) Logical gene ontology annotations (goal): exploring gene ontology annotations with owl. Journal of biomedical semantics 3 Suppl 1: S3.

Conclusion

Le nombre de classes et de relations de Gene Ontology a augmenté de façon monotone entre janvier 2008 et décembre 2012. Considérer indépendemment les trois branches de Gene Ontology (Biological process, Cellular component et Molecular functions) a donné des conclusions similaires mais a mis en lumière différents taux de croissance.

Les métriques de connectivité et celles relatives à la hiérarchie ont fourni d'autres renseignements sur la complexité de l'ontologie. Elles ont révélé différents profils en terme de valeur ainsi qu'au niveau de l'évolution. Les métriques relatives aux graphes comme le degré moyen d'un nœud ont fourni des informations supplémentaires sur la connectivité de l'ontologie. BP et CC ont un degré moyen similaire, supérieur à celui de MF. L'analyse des variations du degré moyen des nœuds a montré que pendant la période d'étude, la connectivité des nœuds de BP a augmenté, alors qu'elle a légèrement diminiée pour CC et est resté stable pour MF. Cela a aussi montré que la baisse concernant CC pouvait être attribuée au fait que le nombre de relations `part of` a moins augmenté que le nombre de classes.

Les métriques relatives à la hiérarchie comme la proportion de feuilles, la profondeur moyenne et la hauteur moyenne des nœuds a donné des informations sur la sémantique. CC et MF ont des proportions de feuilles similaires, une profondeur et hauteur moyennes similaires également. Leurs proportions de feuilles sont supérieures à celles de BP et leurs profondeur et hauteur moyennes inférieures à celles de BP. La proportion de feuilles a diminué pour BP, augmenté pour CC et est restée stable pour MF. La profondeur moyenne des nœuds a augmenté pour BP, est restée plutôt stable pour CC jusqu'à mars 2012 avant de décroître, et est resté plutôt stable pour MF. Ces métriques ont aussi indiqué que la plupart des classes ajoutées à BP n'étaient pas des feuilles mais étaient dans la partie inférieure de la hiérarchie, alors que la plupart des classes ajoutées à CC étaient des nouvelles feuilles ou des sœurs de feuilles existantes, et que la croissance de MF étaient plutôt uniforme. Enfin, les métriques relatives à la hiérarchie seraient capables de distinguer les vraies évolutions de GO des additions et retraits aléatoires de classes et de relations.

Globalement, les résultats ont montré que les trois branches Biological Process, Cellular Component et Molecular Function doivent être considérées séparémment lorsqu'on étudie l'évolution de la complexité de Gene Ontology. Le nombre de classes et de relations a augmenté de façon monotone pour toutes les branches. Nos résultats ont montré que les changements opérés par les personnes en charge de la maintenance de Gene Ontology entre les publications mensuelles ont un impact à la fois sur la taille et la complexité de l'ontologie. La connectivité des nœuds a augmenté de façon monotone pour BP, a baissé globalement pour CC, avec la présence d'extremums locaux, et est restée stable pour MF, avec un profil similaire pour BP et CC comparé à MF. Concernant la hiérarchie, la profondeur moyenne et la hauteur moyenne a augmenté pour BP, diminué pour CC et est restée stable pour MF, CC et MF ayant un profil similaire comparé à BP. Ces résultats indiquent que BP a été sur les quatre dernières années la branche la plus dynamique avec une complexité en augmentation, que CC a subi un ajout de feuilles permettant un niveau d'annotations plus fin, mais diminuant légèrement sa complexité, et que MF a connu une croissance stable et uniforme.

Ces résultats confortent l'idée que les trois branches de Gene Ontology doivent être considérées séparémment lors de mesures de similarité et particularité sémantiques. Ils indiquent également que Gene Ontology est en perpetuelle évolution, chaque branche croissant à son rythme. Cela appelle à recalculer périodiquement nos seuils de similarité et de particularité.

CONCLUSION GÉNÉRALE

La comparaison inter-espèces de voies métaboliques est une problématique importante en biologie. Qualifier et quantifier les caractéristiques communes entre plusieurs espèces ainsi que celles qui les distinguent permet de mieux comprendre le métabolisme de ces espèces. Cela permet également de déterminer si ou dans quelle mesure des résultats obtenus sur une espèce modèle peuvent être transposés à une autre espèce. Cela constitue un enjeu pour la biologie au sens large, avec des répercussion pour la santé humaine aussi bien que pour l'économie. En ce qui concerne le métabolisme des lipides, il existe des pathologies humaines. De plus, la compréhension des mécanismes d'engraissement impacte aussi bien l'économie que le bien-être animal.

Nous avons développé au cours de cette thèse une méthode de comparaison inter-espèces de voies métaboliques. Il faut rappeler que le résultat de cette comparaison dépend grandement de la quantité et de la qualité d'informations disponibles pour chaque espèce que l'on veut comparer. Les données nécessaires sont de trois types : comment s'organise la voie métabolique à comparer chez les deux espèces (quelle est sa *structure* ?), quels sont les produits de gènes qui interviennent à chaque étape de la voie métabolique, et quelle est l'annotation fonctionnelle disponible pour ces produits de gènes. Plus chacun de ces types de données est renseigné, plus le résultat de notre comparaison sera fiable.

La comparaison inter-espèces de voies métaboliques repose sur une ou plusieurs bases de données contenant la succession des réactions chez les espèces à comparer. Des produits de gènes interviennent tout au long de chaque voie métabolique, la plupart en tant qu'enzyme catalysant une réaction. Ces produits de gènes sont annotés par des termes de Gene Ontology, ce qui permet de les comparer entre eux à l'aide d'une mesure de similarité. Comparer des ensembles de termes GO demande une mesure capable de prendre en compte l'héritage qui existe entre ces termes. On parle de mesure de similarité sémantique. Dans notre cas, nous recherchions une mesure qui supportait la comparaison de gènes entre espèces. Cette condition n'est pas respectée par les méthodes basées sur le contenu d'information ("*Information Content*", IC) des termes GO. En effet, l'IC d'un terme dépend de la probabilité qu'il annote un gène. Cette probabilité est calculée par la fréquence à laquelle le terme annote un gène. Il est possible de calculer cette fréquence

sur l'annotation de chaque espèce, menant à autant de valeurs d'IC pour chaque terme qu'il y a d'espèces et empêchant la comparaison inter-espèces. Il est également possible de calculer cette fréquence en cumulant toutes les annotations de toutes les espèces, mais cela conduit à un fort biais en faveur des caractéristiques les mieux connues des espèces les plus étudiées. La comparaison inter-espèce est possible avec les méthodes basées sur les arêtes, puisqu'elles ne dépendent pas d'un corpus d'annotations. Cependant, le fait que la précision des termes GO ne soit pas homogène en fonction de leur profondeur affaiblit la pertinence des résultats obtenus par ces méthodes. La méthode hybride de Wang est celle qui se rapproche le plus d'une méthode basée sur les nœuds sans être dépendante d'un corpus d'annotations. Elle a donc été choisie comme mesure de similarité pour procéder à nos comparaisons inter-espèces.

Comme toutes les mesures de similarité, la mesure de Wang est capable d'attribuer à deux gènes g1 et g2 une similarité haute très proche de celle qu'elle attribue à deux autres gènes g3 et g4 à partir du moment où g1 et g2, comme g3 et g4 ont suffisamment d'annotations en commun, et ce même s'il s'avère que dans une de ces paires de gènes, un gène a en plus des annotations spécifiques qui traduisent des caractéristiques biologiques particulières. Or la comparaison inter-espèces de voies métaboliques se doit de quantifier non seulement la similarité des produits de gènes qui interviennent dans celles-ci, mais également leurs particularités, puisque ce sont principalement celles-ci qui nous intéressent. Nous avions donc besoin d'une mesure de particularité sémantique capable de distinguer des gènes ayant des fonctions particulières même parmi des gènes ayant une forte similarité.

Nous avons donc proposé une mesure de particularité sémantique qui repose sur la notion d'informativité, qui est compatible avec les approches basées sur le contenu d'information aussi bien qu'avec la valeur sémantique de l'approche de Wang. Nous avons démontré l'utilité de la mesure de particularité sémantique, notamment pour identifier et quantifier des caractéristiques propres à un produit de gène comparé à des produits de gènes similaires. Cette mesure ne remplace pas une mesure de similarité, mais devrait être utilisée conjointement à une telle mesure. La mesure de similarité sémantique est symétrique. Ce n'est pas le cas de la mesure de particularité sémantique, puisque la particularité mesurée en comparant A à B est généralement différente de la particularité réciproque. Lorsque l'on compare deux gènes ou deux étapes de voies métaboliques, on obtient donc des profils sous forme de triplets (similarité, particularité, particularité réciproque).

Dans le cadre d'une comparaison inter-espèces, une les configurations de triplets indiquant à la fois une forte similarité et une forte particularité nous permet d'identifer des fonctions propres à une espèce au sein d'un métabolisme qui paraît au premier abord simplement ≪ similaire ≫. Ces cas ne sont pas détectables en utilisant seulement une mesure de similarité sémantique. La comparaison sémantique de produits de gènes repose donc sur l'interprétation des triplets obtenus en utilisant une mesure de similarité et notre mesure de particularité.

Hormis les cas extrèmes, qui sont rarement les plus intéressants, cette interprétation est difficile, faute de savoir à partir de quelle valeur de similarité deux gènes sont similaires, et à partir de quelle valeur de particularité un gène a des fonctions significativement

différentes d'un autre gène. Nous ne disposions donc pas de méthode capable de valider l'interprétation de valeurs de similarité et de particularité de façon à déterminer si deux gènes ou ensembles de gènes sont similaires ou si un gène possède une fonction particulière significative. Cette interprétation étaient jusque là souvent basée soit sur un seuil implicite (on parlait de valeur de similarité « forte » ou « faible ») ou arbitraire (typiquement 0.5, qui représente la moitié de l'intervalle dans lequel se projettent les résultats de la plupart des mesures).

Nous avons donc développé une méthode capable de déterminer un seuil de similarité et un seuil de particularité. La définition de seuils de similarité pour différentes mesures couramment utilisées permet d'identifier les produits de gènes similaires. Le seuil défini pour la méthode de Wang est utile pour identifier des orthologues intervenant dans les voies métaboliques homologues de différentes espèces comme similaires. D'après les résultats obtenus à partir de la base de données HomoloGene, la plupart des orthologues correctement annotés sont similaires. La définition du seuil de particularité permet de savoir si les fonctions spécifiques à un produit de gène lorsqu'on le compare à un produit de gène similaire d'une autre espèce sont anecdotiques ou importantes. Ces seuils nous permettent l'interprétation des résultats obtenus lors de la comparaison inter-espèces systématique de tous les produits de gènes d'une voie métabolique homologue.

Muni de mesures pertinentes et d'une aide à l'interprétation des résultats, nous avons pu procéder à la comparaison d'un métabolisme entre plusieurs espèces. Au travers de l'exemple de la comparaison du métabolisme des lipides chez l'Homme, la souris et la poule, nous avons pu aborder cette problématique sous les trois angles que sont la comparaison de la structure de la voie métabolique, la mesure de la similarité sémantique des produits de gènes présents à chaque étape, et la mesure de leur particularité. Ces trois approches sont complémentaires et leurs résultats peuvent être rassemblés dans un graphe présentant à la fois la structure de la voie métabolique, les points communs et les différences au niveau fonctionnel. L'utilisation d'un seuil de similarité et de particularité a permis de distinguer des cas potentiellement intéressants, qu'ils reflètent une possible réalité biologique (seulement "possible" car sous l'hypothèse d'un monde ouvert) ou une vraisemblable erreur dans une base de données.

LISTE DES TRAVAUX

Publications

Charles Bettembourg, Christian Diot, Anita Burgun et Olivier Dameron, GO2PUB : Querying PubMed with semantic expansion of gene ontology terms, Journal of biomedical semantics **3** :7 (2012), Highly Accessed

Marion Ouédraogo[=], **Charles Bettembourg**[=], Anthony Bretaudeau, Olivier Sallou, Christian Diot, Olivier Demeure et Frederic Lecerf, The Duplicated Genes Database : Identification and Functional Annotation of Co-Localised Duplicated Genes across Genomes, PLoS ONE **7** :11 (2012)

Olivier Dameron, **Charles Bettembourg** et Nolwenn Le Meur, Measuring the Evolution of Ontology Complexity : The Gene Ontology Case Study, PLoS ONE **8** :10 (2013)

Charles Bettembourg, Christian Diot et Olivier Dameron, Semantic particularity measure for functional characterization of gene sets using Gene Ontology, PLoS ONE *in press* (2013)

Charles Bettembourg, Christian Diot et Olivier Dameron, Thresholds of semantic similarity and particularity for gene set functional analysis, *Soumission prochaine* (2013)

Communication orale

Charles Bettembourg, Christian Diot et Olivier Dameron. GO2PUB PubMed Query Tool Based on Semantic Expansion of Gene Ontology Terms, a Lipid Metabolism Case Study, Rennes, France, 03-06 juillet 2012. Journées Ouvertes de la Biologie, de l'Informatique et des Mathématiques.

POSTERS

Charles Bettembourg, Christiant Diot et Olivier Dameron. Comparaison du métabolisme des lipides chez l'humain, la souris et la poule, Limoges, France, 04-06 avril 2011. Séminaire des doctorants de génétique animale.

Charles Bettembourg, Christiant Diot et Olivier Dameron. Comparaison inter-espèces de voies métaboliques, Brest, France, 20 juin avril 2011. Journée des doctorants de l'ifr 140.

Charles Bettembourg, Christiant Diot et Olivier Dameron. Comparaison inter-espèces de voies métaboliques, Paris, France, 28 juin - 1^{er} juillet 2011. Journées Ouvertes de la Biologie, de l'Informatique et des Mathématiques.

SÉMINAIRES INVITÉS

Charles Bettembourg, Christian Diot, Anita Burgun et Olivier Dameron. GO2PUB : a Litterature Search Tool using Gene Ontology, Cambridge, Angleterre, 20 mai 2011. Présentation à l'EBI auprès des groupes UniProt et Dietrich Rebholz-Schumann.

Charles Bettembourg, Christian Diot et Olivier Dameron. Semantic Similarity and Particularity Measures, Jouy-en-Josas, France, 27 juin 2012. Séminaire INRA MIG.

BIBLIOGRAPHIE

Les références dont le numéro de page n'est pas précisé sont citées dans les articles des chapitres 3, 4 et 6.

G. ALTEROVITZ, M. XIANG, D. P. HILL, J. LOMAX, J. LIU, M. CHERKASSKY, J. DREYFUSS, C. MUNGALL, M. A. HARRIS, M. E. DOLAN, J. A. BLAKE et M. F. RAMONI : Ontology engineering. Nature Biotechnology, 28:128–130, 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/16738704.

M. A. ALVAREZ et C. YAN : A graph-based semantic similarity measure for the gene ontology. J Bioinform Comput Biol, 9(6):681–95, déc. 2011. ISSN 0219-7200. URL http://www.ncbi.nlm.nih.gov/pubmed/22084008.

P. ANGULO : Obesity and nonalcoholic fatty liver disease. Nutr Rev, 65(6 Pt 2):S57–63, juin 2007. ISSN 0029-6643. URL http://www.ncbi.nlm.nih.gov/pubmed/17605315. (page 23)

M. E. ARANGUREN, S. BECHHOFER, P. LORD, U. SATTLER et R. STEVENS : Understanding and using the meaning of statements in a bio-ontology : recasting the gene ontology in OWL. BMC bioinformatics, 8:57, 2007.

M. ASHBURNER, C. A. BALL, J. A. BLAKE, D. BOTSTEIN, H. BUTLER, J. M. CHERRY, A. P. DAVIS, K. DOLINSKI, S. S. DWIGHT, J. T. EPPIG, M. A. HARRIS, D. P. HILL, L. ISSEL-TARVER, A. KASARSKIS, S. LEWIS, J. C. MATESE, J. E. RICHARDSON, M. RINGWALD, G. M. RUBIN et G. SHERLOCK : Gene ontology : tool for the unification of biology. the gene ontology consortium. Nat Genet, 25(1):25–9, mai 2000. ISSN 1061-4036. URL http://www.ncbi.nlm.nih.gov/pubmed/10802651. (page 36)

F. AZUAJE, H. WANG, H. ZHENG, O. BODENREIDER et A. CHESNEAU : Predictive integration of gene ontology-driven similarity and functional interactions. 2006. (page 44)

M. BADA, R. STEVENS, C. GOBLE, Y. GIL, M. ASHBURNER, J. A. BLAKE, J. M. CHERRY, M. HARRIS et S. LEWIS : A short study on the success of the gene ontology. Journal of Web Semantics, 1(2):235–240, 2004.

A. BAJPAI, S. DAVULURI, H. HARIDAS, G. KASLIWAL, H. DEEPTI, K. S. SREELAKSHMI, D. S. CHANDRASHEKAR, P. BORA, M. FAROUK, N. CHITTURI, V. SAMUDYATA, K. P. ARUNNEHRU et K. ACHARYA : In search of the right literature search engine(s), 2011. URL http://dx.doi.org/10.1038/npre.2011.2101.3.

E. J. BAKER, J. J. JAY, J. A. BUBIER, M. A. LANGSTON et E. J. CHESLER : Geneweaver : a web-based system for integrative functional genomics. Nucleic Acids Res, 40(Database issue):D1067–76, jan. 2012. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/22080549.

M. BAKER : Genomics : Genomes in three dimensions. Nature, 470(7333):289–94, fév. 2011. ISSN 1476-4687. URL http://www.ncbi.nlm.nih.gov/pubmed/21307943.

D. BAORTO, L. LI et J. J. CIMINO : Practical experience with the maintenance and auditing of a large medical ontology. Journal of biomedical informatics, 42(3):494–503, 2009.

J. B. L. BARD et S. Y. RHEE : Ontologies in biology : design, applications and future challenges. Nat Rev Genet, 5(3):213–22, mars 2004. ISSN 1471-0056. URL http://www.ncbi.nlm.nih.gov/pubmed/14970823. (page 33)

J. D. BARRANS, J. IP, C.-W. LAM, I. L. HWANG, V. J. DZAU et C.-C. LIEW : Chromosomal distribution of the human cardiovascular transcriptome. Genomics, 81(5):519–24, mai 2003. ISSN 0888-7543. URL http://www.ncbi.nlm.nih.gov/pubmed/12706110.

T. BARRETT, D. B. TROUP, S. E. WILHITE, P. LEDOUX, C. EVANGELISTA, I. F. KIM, M. TOMASHEVSKY, K. A. MARSHALL, K. H. PHILLIPPY, P. M. SHERMAN, R. N. MUERTTER, M. HOLKO, O. AYANBULE, A. YEFANOV et A. SOBOLEVA : Ncbi geo : archive for functional genomics data sets–10 years on. Nucleic Acids Res, 39(Database issue):D1005–10, jan. 2011. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/21097893.

R. BARRIOT, D. J. SHERMAN et I. DUTOUR : How to decide which are the most pertinent overly-represented features during gene set enrichment analysis. BMC Bioinformatics, 8:332, 2007. ISSN 1471-2105. URL http://www.ncbi.nlm.nih.gov/pubmed/17848190.

M. BARTON : Mechanisms and therapy of atherosclerosis and its clinical complications. Curr Opin Pharmacol, 13(2):149–53, avr. 2013. ISSN 1471-4973. URL http://www.ncbi.nlm.nih.gov/pubmed/23721738. (page 23)

S. BENABDERRAHMANE, M. SMAIL-TABBONE, O. POCH, A. NAPOLI et M.-D. DEVIGNES : Intelligo : a new vector-based semantic similarity measure including annotation origin. BMC Bioinformatics, 11:588, 2010. ISSN 1471-2105. URL http://www.ncbi.nlm.nih.gov/pubmed/21122125.

W. G. BERGEN et H. J. MERSMANN : Comparative aspects of lipid metabolism : impact on contemporary research and use of animal models. J Nutr, 135(11):2499–502, nov. 2005. ISSN 0022-3166. URL http://www.ncbi.nlm.nih.gov/pubmed/16251600. (pages 19 et 25)

C. BETTEMBOURG, C. DIOT et O. DAMERON : Semantic particularity measure for functional characterization of gene sets using gene ontology. PLoS One (In press), 2013.

J. BHOGAL, A. MACFARLANE et P. SMITH : A review of ontology-based query expansion. Information Processing & Management, 43(4):866–886, juil. 2007.

G. BINDEA, B. MLECNIK, H. HACKL, P. CHAROENTONG, M. TOSOLINI, A. KIRILOVSKY, W.-H. FRIDMAN, F. PAGÈS, Z. TRAJANOSKI et J. GALON : Cluego : a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics, 25(8):1091–3, avr. 2009. ISSN 1367-4811.

G. BLANC et K. H. WOLFE : Functional divergence of duplicated genes formed by polyploidy during arabidopsis evolution. Plant Cell, 16(7):1679–91, juil. 2004. ISSN 1040-4651. URL http://www.ncbi.nlm.nih.gov/pubmed/15208398.

O. BODENREIDER et R. STEVENS : Bio-ontologies : current trends and future directions. Briefings in Bioinformatics, 7(3):256–274, 2006.

N. BOLSHAKOVA, F. AZUAJE et P. CUNNINGHAM : A knowledge-driven approach to cluster validity assessment. Bioinformatics, 21(10):2546–7, mai 2005. ISSN 1367-4803. URL http://www.ncbi.nlm.nih.gov/pubmed/15713738.

S. BORTOLUZZI, L. RAMPOLDI, B. SIMIONATI, R. ZIMBELLO, A. BARBON, F. d'Alessi d'Alessi d'Alessi d'Alessi D'ALESSI, N. TISO, A. PALLAVICINI, S. TOPPO, N. CANNATA, G. VALLE, G. LANFRANCHI et G. A. DANIELI : A comprehensive, high-resolution genomic transcript map of human skeletal muscle. Genome Res, 8(8):817–25, août 1998. ISSN 1088-9051. URL http://www.ncbi.nlm.nih.gov/pubmed/9724327.

E. BOURNEUF, F. HÉRAULT, C. CHICAULT, W. CARRÉ, S. ASSAF, A. MONNIER, S. MOTTIER, S. LAGARRIGUE, M. DOUAIRE, J. MOSSER et C. DIOT : Microarray analysis of differential gene expression in the liver of lean and fat chickens. Gene, 372:162–70, mai 2006. ISSN 0378-1119. URL http://www.ncbi.nlm.nih.gov/pubmed/16513294. (page 24)

E. M. BROWN et H. J. DOWER : Characterization of apolipoproteins from chicken plasma. J Chromatogr, 512:203–12, juil. 1990. ISSN 0021-9673. URL http://www.ncbi.nlm.nih.gov/pubmed/2229228. (page 25)

E. CAMON, M. MAGRANE, D. BARRELL, D. BINNS, W. FLEISCHMANN, P. KERSEY, N. MULDER, T. OINN, J. MASLEN, A. COX et R. APWEILER : The gene ontology annotation (goa) project : implementation of go in swiss-prot, trembl, and interpro. Genome Res, 13 (4):662–72, avr. 2003. ISSN 1088-9051. URL http://www.ncbi.nlm.nih.gov/pubmed/12654719. (page 38)

E. CAMON, M. MAGRANE, D. BARRELL, V. LEE, E. DIMMER, J. MASLEN, D. BINNS, N. HARTE, R. LOPEZ et R. APWEILER : The gene ontology annotation (goa) database : sharing knowledge in uniprot with gene ontology. Nucleic Acids Res, 32(Database issue):D262–6, jan. 2004. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/14681408.

N. CANNATA, E. MERELLI et R. B. ALTMAN : Time to organize the bioinformatics resourceome. PLoS Comput Biol, 1(7):e76, déc. 2005. ISSN 1553-7358. URL http://www.ncbi.nlm.nih.gov/pubmed/16738704.

R. CASPI, T. ALTMAN, K. DREHER, C. A. FULCHER, P. SUBHRAVETI, I. M. KESELER, A. KOTHARI, M. KRUMMENACKER, M. LATENDRESSE, L. A. MUELLER, Q. ONG, S. PALEY, A. PUJAR, A. G. SHEARER, M. TRAVERS, D. WEERASINGHE, P. ZHANG et P. D. KARP : The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucleic Acids Res, 40(Database issue):D742–53, jan. 2012. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/22102576. (page 31)

C. I. CASTILLO-DAVIS, D. L. HARTL et G. ACHAZ : cis-regulatory and protein evolution in orthologous and duplicate genes. Genome Res, 14(8):1530–6, août 2004. ISSN 1088-9051. URL http://www.ncbi.nlm.nih.gov/pubmed/15256508.

W. CEUSTERS : Applying evolutionary terminology auditing to the gene ontology. Journal of biomedical informatics, 42(3):518–529, 2008.

B. CHANG, R. KUSTRA et W. TIAN : Functional-network-based gene set analysis using gene-ontology. PLoS One, 8(2):e55635, 2013. ISSN 1932-6203. URL http://www.ncbi.nlm.nih.gov/pubmed/23418449.

G. CHEN, J. LI et J. WANG : Evaluation of gene ontology semantic similarities on protein interaction datasets. Int J Bioinform Res Appl, 9(2):173–83, 2013. ISSN 1744-5485. URL http://www.ncbi.nlm.nih.gov/pubmed/23467062.

J. CHENG, M. CLINE, J. MARTIN, D. FINKELSTEIN, T. AWAD, D. KULP et M. A. SIANI-ROSE : A knowledge-based clustering algorithm driven by gene ontology. J Biopharm Stat, 14(3):687–700, août 2004. ISSN 1054-3406. URL http://www.ncbi.nlm.nih.gov/pubmed/15468759.

W.-Y. CHUNG, R. ALBERT, I. ALBERT, A. NEKRUTENKO et K. D. MAKOVA : Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network. BMC Bioinformatics, 7:46, 2006. ISSN 1471-2105. URL http://www.ncbi.nlm.nih.gov/pubmed/16441884.

J. J. CIMINO : Auditing the unified medical language system with semantic methods. Journal of the American Medical Informatics Association, 5(1):41–51, 1998.

J. J. CIMINO, T. F. HAYAMIZU, O. BODENREIDER, B. DAVIS, G. A. STAFFORD et M. RINGWALD : The caBIG terminology review process. Journal of biomedical informatics, 42 (3):571–580, 2009.

J. M. CLARK : The epidemiology of nonalcoholic fatty liver disease in adults. J Clin Gastroenterol, 40 Suppl 1:S5–10, mars 2006. ISSN 0192-0790. URL http://www.ncbi.nlm.nih.gov/pubmed/16540768. (pages 23 et 24)

J. M. CLARK et A. M. DIEHL : Nonalcoholic fatty liver disease : an underrecognized cause of cryptogenic cirrhosis. JAMA, 289(22):3000–4, juin 2003. ISSN 1538-3598. URL http://www.ncbi.nlm.nih.gov/pubmed/12799409. (page 23)

J. COHEN : A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1):37–46, 1960.

F. M. COUTO, M. J. SILVA et P. COUTINHO : Semantic similarity over the gene ontology : family correlation and selecting disjunctive ancestors. In O. HERZOG, H.-J. SCHEK, N. FUHR, A. CHOWDHURY et W. TEIKEN, éds : CIKM, p. 343–344. ACM, 2005. ISBN 1-59593-140-6.

F. M. COUTO, M. J. SILVA et P. M. COUTINHO : Measuring semantic similarity between gene ontology terms. Data & Knowledge Engineering, 61:137–152, April 2007. (pages 44, 46 et 50)

D. CROFT, G. O'KELLY, G. WU, R. HAW, M. GILLESPIE, L. MATTHEWS, M. CAUDY, P. GARAPATI, G. GOPINATH, B. JASSAL, S. JUPE, I. KALATSKAYA, S. MAHAJAN, B. MAY, N. NDEGWA, E. SCHMIDT, V. SHAMOVSKY, C. YUNG, E. BIRNEY, H. HERMJAKOB, P. D'EUSTACHIO et L. STEIN : Reactome : a database of reactions, pathways and biological processes. Nucleic Acids Res, 39(Database issue):D691–7, jan. 2011. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/21067998. (page 30)

O. DAMERON, C. BETTEMBOURG et N. LE MEUR : Measuring the evolution of ontology complexity : the gene ontology case study. PLoS One, 8(10):e75993, 2013. ISSN 1932-6203. URL http://www.ncbi.nlm.nih.gov/pubmed/24146805.

S. DAVAL, S. LAGARRIGUE et M. DOUAIRE : Messenger rna levels and transcription rates of hepatic lipogenesis genes in genetically lean and fat chickens. Genet Sel Evol, 32 (5):521–31, 2000. ISSN 0999-193X. URL http://www.ncbi.nlm.nih.gov/pubmed/14736380. (pages 24 et 25)

S. de CORONADO, L. W. WRIGHT, G. FRAGOSO, M. W. HABER, E. A. HAHN-DANTONA, F. W. HARTEL, S. L. QUAN, T. SAFRAN, N. THOMAS et L. WHITEMAN : The nci thesaurus quality assurance life cycle. Journal of biomedical informatics, 42(3):530–539, 2009.

B. DESVERGNE, L. MICHALIK et W. WAHLI : Transcriptional regulation of metabolism. Physiol Rev, 86(2):465–514, avr. 2006. ISSN 0031-9333. URL http://www.ncbi.nlm.nih.gov/pubmed/16601267.

N. DÍAZ-DÍAZ et J. S. AGUILAR-RUIZ : Go-based functional dissimilarity of gene sets. BMC Bioinformatics, 12:360, 2011. ISSN 1471-2105. URL http://www.ncbi.nlm.nih.gov/pubmed/21884611.

E. W. DIJKSTRA : A note on two problems in connexion with graphs. Numer. Math., 1:269–271, 1959. (page 44)

E. C. DIMMER, R. P. HUNTLEY, Y. ALAM-FARUQUE, T. SAWFORD, C. O'DONOVAN, M. J. MARTIN, B. BELY, P. BROWNE, W. MUN CHAN, R. EBERHARDT, M. GARDNER, K. LAIHO, D. LEGGE, M. MAGRANE, K. PICHLER, D. POGGIOLI, H. SEHRA, A. AUCHINCLOSS, K. AXELSEN, M.-C. BLATTER, E. BOUTET, S. BRACONI-QUINTAJE, L. BREUZA, A. BRIDGE, E. COUDERT, A. ESTREICHER, L. FAMIGLIETTI, S. FERRO-ROJAS, M. FEUERMANN, A. GOS, N. GRUAZ-GUMOWSKI, U. HINZ, C. HULO, J. JAMES, S. JIMENEZ, F. JUNGO, G. KELLER, P. LEMERCIER, D. LIEBERHERR, P. MASSON, M. MOINAT, I. PEDRUZZI, S. POUX, C. RIVOIRE, B. ROECHERT, M. SCHNEIDER, A. STUTZ, S. SUNDARAM, M. TOGNOLLI, L. BOUGUELERET, G. ARGOUD-PUY, I. CUSIN, P. DUEK-ROGGLI, I. XENARIOS et R. APWEILER : The uniprot-go annotation database in 2011. Nucleic Acids Res, 40(Database issue):D565–70, jan. 2012. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/22123736. (pages 125 et 132)

A. DOMS et M. SCHROEDER : Gopubmed : exploring pubmed with the gene ontology. Nucleic Acids Res., 33:W783–W786, 2005.

L. DURET, D. MOUCHIROUD et M. GOUY : Hovergen : a database of homologous vertebrate genes. Nucleic Acids Res, 22(12):2360–5, juin 1994. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pubmed/8036164.

T. EITER, G. IANNI, A. POLLERES, R. SCHINDLAUER et H. TOMPITS : Reasoning with rules and ontologies. LNCS 4126, 2006. (page 34)

E. FAHY, S. SUBRAMANIAM, R. C. MURPHY, M. NISHIJIMA, C. R. H. RAETZ, T. SHIMIZU, F. SPENER, G. van MEER, M. J. O. WAKELAM et E. A. DENNIS : Update of the lipid maps comprehensive classification system for lipids. J Lipid Res, 50 Suppl:S9–14, avr. 2009. ISSN 0022-2275. URL http://www.ncbi.nlm.nih.gov/pubmed/19098281. (page 18)

D. FARIA, A. SCHLICKER, C. PESQUITA, H. BASTOS, A. E. N. FERREIRA, M. ALBRECHT et A. O. FALCAO : Mining GO annotations for improving annotation consistency. PloS one, 7(7):e40519, 2012.

D. FARRÉ et M. M. ALBÀ : Heterogeneous patterns of gene-expression diversification in mammalian gene duplicates. Mol Biol Evol, 27(2):325–35, fév. 2010. ISSN 1537-1719. URL http://www.ncbi.nlm.nih.gov/pubmed/19822635.

C. FELLBAUM : WordNet : An Electronic Lexical Database. MIT Press, 1998.

O. FILANGI, Y. BEAUSSE, A. ASSI, L. LEGRAND, J.-M. LARRÉ, V. MARTIN, O. COLLIN, C. CARON, H. LEROY et D. ALLOUCHE : Biomaj : a flexible framework for databanks synchronization and processing. Bioinformatics, 24(16):1823–5, août 2008. ISSN 1367-4811. URL http://www.ncbi.nlm.nih.gov/pubmed/18593718.

R. FRASER, V. R. HESLOP, F. E. MURRAY et W. A. DAY : Ultrastructural studies of the portal transport of fat in chickens. Br J Exp Pathol, 67(6):783–91, déc. 1986. ISSN 0007-1021. URL http://www.ncbi.nlm.nih.gov/pubmed/3801295. (page 24)

R. FRIEDMAN et A. L. HUGHES : Gene duplication and the structure of eukaryotic genomes. Genome Res, 11(3):373–81, mars 2001. ISSN 1088-9051. URL http://www.ncbi.nlm.nih.gov/pubmed/11230161.

Y. FUKUOKA, H. INAOKA et I. S. KOHANE : Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. BMC Genomics, 5(1):4, jan. 2004. ISSN 1471-2164. URL http://www.ncbi.nlm.nih.gov/pubmed/14718066.

A. GADDI, A. F. G. CICERO, F. O. ODOO, A. A. POLI, R. PAOLETTI et ATHEROSCLEROSIS AND METABOLIC DISEASES STUDY GROUP : Practical guidelines for familial combined hyperlipidemia diagnosis : an up-date. Vasc Health Risk Manag, 3(6):877–86, 2007. ISSN 1176-6344. URL http://www.ncbi.nlm.nih.gov/pubmed/18200807. (page 23)

M. GAN, X. DOU et R. JIANG : From ontology to semantic similarity : calculation of ontology-based semantic similarity. ScientificWorldJournal, 2013:793091, 2013. ISSN 1537-744X. URL http://www.ncbi.nlm.nih.gov/pubmed/23533360.

E. W. GANKO, B. C. MEYERS et T. J. VISION : Divergence in expression between duplicated genes in arabidopsis. Mol Biol Evol, 24(10):2298–309, oct. 2007. ISSN 0737-4038. URL http://www.ncbi.nlm.nih.gov/pubmed/17670808.

GENOUEST : Genouest xref server : a webservice dedicated to cross-references., 2012.

J. GILLIS et P. PAVLIDIS : Assessing identity, redundancy and confounds in gene ontology annotations over time. Bioinformatics, 29(4):476–82, fév. 2013. ISSN 1367-4811. URL http://www.ncbi.nlm.nih.gov/pubmed/23297035.

C. GOLBREICH, M. HORRIDGE, I. HORROCKS, B. MOTIK et R. SHEARER : OBO and OWL : Leveraging semantic web technologies for the life sciences. In Proceedings of the 6th International Semantic Web Conference (ISWC 2007), vol. 4825 de Lecture Notes in Computer Science, p. 169–182, 2007.

H. D. GRIFFIN, K. GUO, D. WINDSOR et S. C. BUTTERWITH : Adipose tissue lipogenesis and fat deposition in leaner broiler chickens. J Nutr, 122(2):363–8, fév. 1992. ISSN 0022-3166. URL http://www.ncbi.nlm.nih.gov/pubmed/1732477. (page 24)

N. GRIFFON, W. CHEBIL, L. ROLLIN, G. KERDELHUE, B. THIRION, J.-F. GEHANNO et S. J. DARMONI : Performance evaluation of unified medical language system®'s synonyms expansion to query pubmed. BMC Med Inform Decis Mak, 12:12, 2012. ISSN 1472-6947.

A. GROSS, M. HARTUNG, K. PRÜFER, J. KELSO et E. RAHM : Impact of ontology evolution on functional analyses. Bioinformatics (Oxford, England), 28(20):2671–2677, 2012.

S. GROSSMANN, S. BAUER, P. N. ROBINSON et M. VINGRON : Improved detection of over-representation of gene-ontology annotations with parent child analysis. Bioinformatics, 23(22):3024–31, nov. 2007. ISSN 1367-4811. URL http://www.ncbi.nlm.nih.gov/pubmed/17848398.

H. H. GU, D. WEI, J. L. V. MEJINO et G. ELHANAN : Relationship auditing of the fma ontology. Journal of biomedical informatics, 42(3):550–557, 2009.

Z. GU, D. NICOLAE, H. H.-S. LU et W. H. LI : Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet, 18(12):609–13, déc. 2002. ISSN 0168-9525. URL http://www.ncbi.nlm.nih.gov/pubmed/12446139.

Z. GU, S. A. RIFKIN, K. P. WHITE et W.-H. LI : Duplicate genes increase gene expression diversity within and between species. Nat Genet, 36(6):577–9, juin 2004. ISSN 1061-4036. URL http://www.ncbi.nlm.nih.gov/pubmed/15122255.

Z. GU, L. M. STEINMETZ, X. GU, C. SCHARFE, R. W. DAVIS et W.-H. LI : Role of duplicate genes in genetic robustness against null mutations. Nature, 421(6918):63–6, jan. 2003. ISSN 0028-0836. URL http://www.ncbi.nlm.nih.gov/pubmed/12511954.

N. GUARINO et C. WELTY : Evaluating ontological decisions with Ontoclean. In Communications of the ACM, vol. 45, p. 61–65, 2002.

R. J. HANSEN et R. L. WALZEM : Avian fatty liver hemorrhagic syndrome : a comparative review. Adv Vet Sci Comp Med, 37:451–68, 1993. ISSN 0065-3519. URL http://www.ncbi.nlm.nih.gov/pubmed/8273524. (page 25)

M. A. HARRIS et G. O. CONSORTIUM : The gene ontology (go) database and informatics resource. Nucleic Acids Res., 1:D258–D261, 2004.

M. HARTUNG, A. GROSS et E. RAHM : CODEX : exploration of semantic changes between ontology versions. Bioinformatics (Oxford, England), 28(6):895–896, 2012.

M. HARTUNG, T. KIRSTEN, A. GROSS et E. RAHM : Onex : Exploring changes in life science ontologies. BMC bioinformatics, 10:250, 2009.

M. HARTUNG, K. TORALF et E. RAHM : Analyzing the evolution of life science ontologies and mappings. In 5th Intl. Workshop on Data Integration in the Life Sciences (DILS) 2008, LNCS 5109, 2008.

T. HAWKINS, M. CHITALE et D. KIHARA : Functional enrichment analyses and construction of functional similarity networks with high confidence function prediction by pfp. BMC Bioinformatics, 11:265, 2010. ISSN 1471-2105. URL http://www.ncbi.nlm.nih.gov/pubmed/20482861.

S. B. HEDGES : The origin and evolution of model organisms. Nat Rev Genet, 3(11):838–49, nov. 2002. ISSN 1471-0056. URL http://www.ncbi.nlm.nih.gov/pubmed/12415314.

D. HERMIER : Lipoprotein metabolism and fattening in poultry. J Nutr, 127(5 Suppl):805S–808S, mai 1997. ISSN 0022-3166. URL http://www.ncbi.nlm.nih.gov/pubmed/9164241. (pages 24, 25 et 26)

D. P. HILL, B. SMITH, M. S. MCANDREWS-HILL et J. A. BLAKE : Gene ontology annotations : what they mean and where they come from. BMC Bioinformatics, 9 Suppl 5:S2, 2008. ISSN 1471-2105. URL http://www.ncbi.nlm.nih.gov/pubmed/18460184. (page 38)

R. HOEHNDORF, M. DUMONTIER et G. V. GKOUTOS : Evaluation of research in biomedical ontologies. Briefings in bioinformatics, 2012. In press.

T.-L. HSIAO et D. VITKUP : Role of duplicate genes in robustness against deleterious human mutations. PLoS Genet, 4(3):e1000014, mars 2008. ISSN 1553-7404. URL http://www.ncbi.nlm.nih.gov/pubmed/18369440.

D. W. HUANG, B. T. SHERMAN et R. A. LEMPICKI : Bioinformatics enrichment tools : paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res, 37 (1):1–13, jan. 2009. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/19033363.

T. HULSEN, J. de VLIEG et W. ALKEMA : Biovenn - a web application for the comparison and visualization of biological lists using area-proportional venn diagrams. BMC Genomics, 9(1):488, 2008.

L. HUMINIECKI et K. H. WOLFE : Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res, 14 (10A):1870–9, oct. 2004. ISSN 1088-9051. URL http://www.ncbi.nlm.nih.gov/pubmed/15466287.

J.-H. HUNG, T.-H. YANG, Z. HU, Z. WENG et C. DELISI : Gene set enrichment analysis : performance evaluation and usage guidelines. Brief Bioinform, 13(3):281–91, mai 2012. ISSN 1477-4054. URL http://www.ncbi.nlm.nih.gov/pubmed/21900207.

M. M. HUSSAIN, R. K. KANCHA, Z. ZHOU, J. LUCHOOMUN, H. ZU et A. BAKILLAH : Chylomicron assembly and catabolism : role of apolipoproteins and receptors. Biochim Biophys Acta, 1300(3):151–70, mai 1996. ISSN 0006-3002. URL http://www.ncbi.nlm.nih.gov/pubmed/8679680. (page 19)

J. JIANG et D. CONRATH : Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the International Conference Research on Computational Linguistics (ROCLING), Taiwan, 1997. (page 45)

B. JIN et X. LU : Identifying informative subsets of the gene ontology with information bottleneck methods. Bioinformatics, 26(19):2445–51, oct. 2010. ISSN 1367-4811. URL http://www.ncbi.nlm.nih.gov/pubmed/20702400.

Z. JUN, A. SILVESCU et V. HONAVAR : Ontology-driven induction of decision trees at multiple levels of abstraction. In S. KOENIG et R. C. HOLTE, éds : SARA, vol. 2371 de Lecture Notes in Computer Science, p. 316–323. Springer, 2002. ISBN 3-540-43941-2. (page 34)

S. JUPP, R. STEVENS et R. HOEHNDORF : Logical gene ontology annotations (goal) : exploring gene ontology annotations with owl. Journal of biomedical semantics, 3 Suppl 1:S3, 2012.

S. KADAUKE et G. A. BLOBEL : Chromatin loops in gene regulation. Biochim Biophys Acta, 1789(1):17–25, jan. 2009. ISSN 0006-3002. URL http://www.ncbi.nlm.nih.gov/pubmed/18675948.

M. KANEHISA et S. GOTO : Kegg : kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 28(1):27–30, jan. 2000. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pubmed/10592173. (page 32)

P. KHATRI et S. DRAGHICI : Ontological analysis of gene expression data : current tools, limitations, and open problems. Bioinformatics (Oxford, England), 21(18):3587–3595, 2005.

T. KIRSTEN, A. GROSS, M. HARTUNG et E. RAHM : Gomma : A component-based infrastructure for managing and analyzing life science ontologies and their evolution. Journal of biomedical semantics, 2(1):6, 2011.

S. KLIE, M. MUTWIL, S. PERSSON et Z. NIKOLOSKI : Inferring gene functions through dissection of relevance networks : interleaving the intra- and inter-species views. Mol Biosyst, 8(9):2233–41, sept. 2012. ISSN 1742-2051. URL http://www.ncbi.nlm.nih.gov/pubmed/22744313.

M. S. KO, T. A. THREAT, X. WANG, J. H. HORTON, Y. CUI, X. WANG, E. PRYOR, J. PARIS, J. WELLS-SMITH, J. R. KITCHEN, L. B. ROWE, J. EPPIG, T. SATOH, L. BRANT, H. FUJIWARA, S. YOTSUMOTO et H. NAKASHIMA : Genome-wide mapping of unselected transcripts from extraembryonic tissue of 7.5-day mouse embryos reveals enrichment in the t-complex and under-representation on the x chromosome. Hum Mol Genet, 7(12):1967–78, nov. 1998. ISSN 0964-6906. URL http://www.ncbi.nlm.nih.gov/pubmed/9811942.

J. KÖHLER, K. MUNN, A. RÜEGG, A. SKUSA et B. SMITH : Quality control for terms and definitions in ontologies and taxonomies. BMC Bioinformatics, 7:212, 2006.

S. KÖHLER, S. BAUER, C. J. MUNGALL, G. CARLETTI, C. L. SMITH, P. SCHOFIELD, G. V. GKOUTOS et P. N. ROBINSON : Improving ontologies by automatic reasoning and evaluation of logical definitions. BMC bioinformatics, 12(1):418, 2011.

S. KRISHNAN, L. CLEMENTI, J. REN, P. PAPADOPOULOS et W. LI : Design and evaluation of opal2 : A toolkit for scientific software as a service. IEEE Computer Society, p. 709–716, 2009.

L. KULIK, M. DUCKHAM et M. J. EGENHOFER : Ontology-driven map generalization. J. Vis. Lang. Comput., 16(3):245–267, 2005. (page 34)

R. KUSTRA et A. ZAGDANSKI : Incorporating gene ontology in clustering gene expression data. In CBMS, p. 555–563. IEEE Computer Society, 2006.

J. R. LANDIS et G. G. KOCH : The measurement of observer agreement for categorical data. Biometrics, 33(1):159–74, 1977.

S. LEONELLI, A. D. DIEHL, K. R. CHRISTIE, M. A. HARRIS et J. LOMAX : How the gene ontology evolves. BMC bioinformatics, 12(1):325, 2011.

M. J. LERCHER, T. BLUMENTHAL et L. D. HURST : Coexpression of neighboring genes in caenorhabditis elegans is mostly due to operons and duplicate genes. Genome Res, 13 (2):238–43, fév. 2003. ISSN 1088-9051. URL http://www.ncbi.nlm.nih.gov/pubmed/12566401.

M. J. LERCHER, A. O. URRUTIA et L. D. HURST : Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet, 31(2):180–3, juin 2002. ISSN 1061-4036. URL http://www.ncbi.nlm.nih.gov/pubmed/11992122.

Q. LI, B. T. K. LEE et L. ZHANG : Genome-scale analysis of positional clustering of mouse testis-specific genes. BMC Genomics, 6:7, 2005a. ISSN 1471-2164. URL http://www.ncbi.nlm.nih.gov/pubmed/15656914.

W. H. LI, Z. GU, H. WANG et A. NEKRUTENKO : Evolutionary analyses of the human genome. Nature, 409(6822):847–9, fév. 2001. ISSN 0028-0836. URL http://www.ncbi.nlm.nih.gov/pubmed/11237007.

W.-H. LI, J. YANG et X. GU : Expression divergence between duplicate genes. Trends Genet, 21(11):602–7, nov. 2005b. ISSN 0168-9525. URL http://www.ncbi.nlm.nih.gov/pubmed/16140417.

D. LIN : An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning, p. 296–304, 1998. (page 45)

S. LOGUERCIO, E. L. CLARKE, B. M. GOOD et A. I. SU : A task-based approach for large-scale evaluation of the gene ontology. In IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology, HISB 2012, p. 144, 2012.

P. W. LORD, R. D. STEVENS, A. BRASS et C. A. GOBLE : Semantic similarity measures as tools for exploring the gene ontology. Pac Symp Biocomput, p. 601–12, 2003. ISSN 2335-6936. URL http://www.ncbi.nlm.nih.gov/pubmed/12603061. (page 44)

Z. LU : Pubmed and beyond : a survey of web tools for searching biomedical literature. Database (Oxford), p. baq036, 2011.

Z. LU, W. KIM et W. J. WILBUR : Evaluation of query expansion using mesh in pubmed. Inf Retr Boston, 12(1):69–80, 2009.

S. MAERE, K. HEYMANS et M. KUIPER : Bingo : a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics, 21(16):3448–9, août 2005. ISSN 1367-4803. URL http://www.ncbi.nlm.nih.gov/pubmed/15972284.

P. MAGNUS et R. BEAGLEHOLE : The real contribution of the major risk factors to the coronary epidemics : time to end the "only-50%" myth. Arch Intern Med, 161(22):2657–60, déc. 2001. ISSN 0003-9926. URL http://www.ncbi.nlm.nih.gov/pubmed/11732929. (page 23)

K. D. MAKOVA et W.-H. LI : Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res, 13(7):1638–45, juil. 2003. ISSN 1088-9051. URL http://www.ncbi.nlm.nih.gov/pubmed/12840042.

J. MALONE et R. STEVENS : Measuring the level of activity in community built bio-ontologies. Journal of biomedical informatics, 46(1):5–14, 2012.

S. MATOS, J. P. ARRAIS, J. MAIA-RODRIGUES et J. L. OLIVEIRA : Concept-based query expansion for retrieving gene related publications from medline. BMC Bioinformatics, 11:212, 2010.

G. K. MAZANDU et N. J. MULDER : A topology-based metric for measuring term similarity in the gene ontology. Adv Bioinformatics, 2012:975783, 2012. ISSN 1687-8035. URL http://www.ncbi.nlm.nih.gov/pubmed/22666244. (page 43)

T. F. MEEHAN, A. M. MASCI, A. ABDULLA, L. G. COWELL, J. A. BLAKE, C. J. MUNGALL et A. D. DIEHL : Logical development of the cell ontology. BMC bioinformatics, 12(1):6, 2011.

H. MI, B. LAZAREVA-ULITSKY, R. LOO, A. KEJARIWAL, J. VANDERGRIFF, S. RABKIN, N. GUO, A. MURUGANUJAN, O. DOREMIEUX, M. J. CAMPBELL, H. KITANO et P. D. THOMAS : The panther database of protein families, subfamilies, functions and pathways. Nucleic Acids Res, 33(Database issue):D284–8, jan. 2005. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/15608197.

L. MICHALIK, B. DESVERGNE, C. DREYER, M. GAVILLET, R. N. LAURINI et W. WAHLI : Ppar expression and function during vertebrate development. Int J Dev Biol, 46(1):105–14, jan. 2002. ISSN 0214-6282. URL http://www.ncbi.nlm.nih.gov/pubmed/11902671.

G. MILLER : Wordnet : A lexical database for english. Communications of the ACM, 38 (1):39–41, 1995.

S. MINAGAWA, K. NAKABAYASHI, M. FUJII, S. W. SCHERER et D. AYUSAWA : Functional and chromosomal clustering of genes responsive to 5-bromodeoxyuridine in human cells. Exp Gerontol, 39(7):1069–78, juil. 2004. ISSN 0531-5565. URL http://www.ncbi.nlm.nih.gov/pubmed/15236766.

J. B. MOORE : Non-alcoholic fatty liver disease : the hepatic consequence of obesity and the metabolic syndrome. Proc Nutr Soc, 69(2):211–20, mai 2010. ISSN 1475-2719. URL http://www.ncbi.nlm.nih.gov/pubmed/20158939. (page 23)

M. MUIN, P. FONTELO, F. LIU et M. ACKERMAN : Slim : an alternative web interface for medline/pubmed searches - a preliminary study. BMC Med Inform Decis Mak., 5:37, 2005.

C. J. MUNGALL, M. BADA, T. Z. BERARDINI, J. DEEGAN, A. IRELAND, M. A. HARRIS, D. P. HILL et J. LOMAX : Cross-product extensions of the gene ontology. Journal of biomedical informatics, 44(1):80–86, 2010.

R. NAVIGLI et P. VELARDI : An analysis of ontology-based query expansion strategies. Proceedings of the 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia, p. 42–49, 2003.

NCBI RESOURCE COORDINATORS : Database resources of the national center for biotechnology information. Nucleic Acids Res, 41(Database issue):D8–D20, jan. 2013. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/23193264.

Y. K. NG, W. WU et L. ZHANG : Positive correlation between gene coexpression and positional clustering in the zebrafish genome. BMC Genomics, 10:42, 2009. ISSN 1471-2164. URL http://www.ncbi.nlm.nih.gov/pubmed/19159490.

Y. NIIMURA et M. NEI : Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates. J Hum Genet, 51(6):505–17, 2006. ISSN 1434-5161. URL http://www.ncbi.nlm.nih.gov/pubmed/16607462.

M. F. OCHS, A. J. PETERSON, A. KOSSENKOV et G. BIDAUT : Incorporation of gene ontology annotations to enhance microarray data analysis. Methods Mol Biol, 377:243–54, 2007. ISSN 1064-3745. URL http://www.ncbi.nlm.nih.gov/pubmed/17634621.

R. M. OTHMAN, S. DERIS et R. M. ILLIAS : A genetic similarity algorithm for searching the gene ontology terms and annotating anonymous protein sequences. J Biomed Inform, 41(1):65–81, fév. 2008. ISSN 1532-0480. URL http://www.ncbi.nlm.nih.gov/pubmed/17681495. (page 49)

S. M. PALEY et P. D. KARP : Evaluation of computational metabolic-pathway predictions for helicobacter pylori. Bioinformatics, 18(5):715–24, mai 2002. ISSN 1367-4803. URL http://www.ncbi.nlm.nih.gov/pubmed/12050068. (page 32)

D. PAPANDREOU, I. ROUSSO et I. MAVROMICHALIS : Update on non-alcoholic fatty liver disease in children. Clin Nutr, 26(4):409–15, août 2007. ISSN 0261-5614. URL http://www.ncbi.nlm.nih.gov/pubmed/17449148. (page 23)

J. C. PARK, T. eun KIM et J. PARK : Monitoring the evolutionary aspect of the gene ontology to enhance predictability and usability. BMC bioinformatics, 9 Suppl 3:S7, 2008.

Y. R. PARK, J. KIM, H. W. LEE, Y. J. YOON et J. H. KIM : GOChase-II : correcting semantic inconsistencies from gene ontology-based annotations for gene products. BMC bioinformatics, 12 Suppl 1:S40, 2011.

V. PEKAR et S. STAAB : Taxonomy learning - factoring the structure of a taxonomy into a semantic classification decision. In COLING, 2002. (page 45)

C. PESQUITA et F. M. COUTO : Where GO is going and what it means for ontology extension. In International Conference on Biomedical Ontology ICBO, 2011.

C. PESQUITA, D. FARIA, H. BASTOS, A. E. FERREIRA, O. FALCAON ANDRÉ et F. M. COUTO : Metrics for go based protein semantic similarity : a systematic evaluation. BMC Bioinformatics, 9(Suppl 5):S4, 2008.

C. PESQUITA, D. FARIA, A. O. FALCÃO, P. LORD et F. M. COUTO : Semantic similarity in biomedical ontologies. PLoS Comput Biol, 5(7):e1000443, juil. 2009. ISSN 1553-7358. URL http://www.ncbi.nlm.nih.gov/pubmed/19649320. (pages 43 et 50)

A. R. PICO, T. KELDER, M. P. van IERSEL, K. HANSPERS, B. R. CONKLIN et C. EVELO : Wikipathways : pathway editing for the people. PLoS Biol, 6(7):e184, juil. 2008. ISSN 1545-7885. URL http://www.ncbi.nlm.nih.gov/pubmed/18651794. (page 32)

F. PITEL, T. FARAUT, G. BRUNEAU et P. MONGET : Is there a leptin gene in the chicken genome ? lessons from phylogenetics, bioinformatics and genomics. Gen Comp Endocrinol, 167(1):1–5, mai 2010. ISSN 1095-6840. URL http://www.ncbi.nlm.nih.gov/pubmed/19854194. (page 25)

C. R. PRIMMER, S. PAPAKOSTAS, E. H. LEDER, M. J. DAVIS et M. A. RAGAN : Annotated genes and nonannotated genomes : cross-species use of gene ontology in ecology and evolution research. Mol Ecol, juin 2013. ISSN 1365-294X. URL http://www.ncbi.nlm.nih.gov/pubmed/23763602.

M. PUNTA, P. C. COGGILL, R. Y. EBERHARDT, J. MISTRY, J. TATE, C. BOURSNELL, N. PANG, K. FORSLUND, G. CERIC, J. CLEMENTS, A. HEGER, L. HOLM, E. L. L. SONNHAMMER, S. R. EDDY, A. BATEMAN et R. D. FINN : The pfam protein families database. Nucleic Acids Res, 40(Database issue):D290–301, jan. 2012. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/22127870.

A. PURMANN, J. TOEDLING, M. SCHUELER, P. CARNINCI, H. LEHRACH, Y. HAYASHIZAKI, W. HUBER et S. SPERLING : Genomic organization of transcriptomes in mammals : Coregulation and cofunctionality. Genomics, 89(5):580–7, mai 2007. ISSN 0888-7543. URL http://www.ncbi.nlm.nih.gov/pubmed/17369017.

Y. QIAN et J.-G. FAN : Obesity, fatty liver and liver cancer. Hepatobiliary Pancreat Dis Int, 4 (2):173–7, mai 2005. ISSN 1499-3872. URL http://www.ncbi.nlm.nih.gov/pubmed/15908310. (page 23)

R. RADA, H. MILI, E. BICKNELL et M. BLETTNER : Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man, and Cybernetics, 19(1):17–30, 1989. ISSN 0018-9472. (page 44)

D. REBHOLZ-SCHUHMANN, H. KIRSCH, M. ARREGUI, S. GAUDAN, M. RIETHOVEN et P. STOEHR : Ebimed–text crunching to gather facts for proteins from medline. Bioinformatics, 23(2):e237–44, jan. 2007. ISSN 1367-4811.

A. L. RECTOR, S. BRANDT et T. SCHNEIDER : Getting the foot out of the pelvis : modeling problems affecting use of snomed ct hierarchies in practical applications. Journal of the American Medical Informatics Association : JAMIA, 18(4):432–440, 2011.

J. K. REDDY et M. S. RAO : Lipid metabolism and liver inflammation. ii. fatty liver disease and fatty acid oxidation. Am J Physiol Gastrointest Liver Physiol, 290(5):G852–8, mai 2006. ISSN 0193-1857. URL http://www.ncbi.nlm.nih.gov/pubmed/16603729. (pages 23 et 24)

X.-Y. REN, M. W. E. J. FIERS, W. J. STIEKEMA et J.-P. NAP : Local coexpression domains of two to four genes in the genome of arabidopsis. Plant Physiol, 138(2):923–34, juin 2005. ISSN 0032-0889. URL http://www.ncbi.nlm.nih.gov/pubmed/15923337.

P. RESNIK : Semantic similarity in a taxonomy : An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence, 11(11):95–130, 1999. URL http://scholar.google.com/scholar?hl=de&lr=&cluster=6503302567385908046. (page 45)

S. Y. RHEE, V. WOOD, K. DOLINSKI et S. DRAGHICI : Use and misuse of the gene ontology annotations. Nat Rev Genet, 9(7):509–15, juil. 2008. ISSN 1471-0064. URL http://www.ncbi.nlm.nih.gov/pubmed/18475267. (pages 38 et 49)

P. ROMERO, J. WAGG, M. L. GREEN, D. KAISER, M. KRUMMENACKER et P. D. KARP : Computational prediction of human metabolic pathways from the complete human genome. Genome Biol, 6(1):R2, 2005. ISSN 1465-6914. URL http://www.ncbi.nlm.nih.gov/pubmed/15642094.

B. ROST : Twilight zone of protein sequence alignments. Protein Eng, 12(2):85–94, fév. 1999. ISSN 0269-2139. URL http://www.ncbi.nlm.nih.gov/pubmed/10195279.

A. SCHLICKER, F. S. DOMINGUES, J. RAHNENFÜHRER et T. LENGAUER : A new measure for functional similarity of gene products based on gene ontology. BMC Bioinformatics, 7:302, 2006. ISSN 1471-2105. URL http://www.ncbi.nlm.nih.gov/pubmed/16776819. (page 46)

J. L. SEVILLA, V. SEGURA, A. PODHORSKI, E. GURUCEAGA, J. M. MATO, L. A. MARTINEZ-CRUZ, F. J. CORRALES et A. RUBIO : Correlation between gene expression and go semantic similarity. IEEE/ACM Trans Comput Biol Bioinform, 2(4):330–8, 2005. ISSN 1545-5963. URL http://www.ncbi.nlm.nih.gov/pubmed/17044170. (page 44)

Y. SHAHAR, H. CHEN, D. P. STITES, L. V. BASSO, H. KAIZER, D. M. WILSON et M. A. MUSEN : Semi-automated entry of clinical temporal-abstraction knowledge. J Am Med Inform Assoc, 6(6):494–511, 1999. ISSN 1067-5027. URL http://www.ncbi.nlm.nih.gov/pubmed/10579607. (page 34)

C. E. SHANNON : A mathematical theory of communication. Bell system technical journal, 27, 1948.

K. SHCHEKOTYKHIN, G. FRIEDRICH, P. FLEISS et P. RODLER : Interactive ontology debugging : Two query strategies for efficient fault localization. Web semantics (Online), 12-13 (C):88–103, 2012.

B. SHEEHAN, A. QUIGLEY, B. GAUDIN et S. DOBSON : A relation based measure of semantic similarity for gene ontology annotations. BMC Bioinformatics, 9:468, 2008. ISSN 1471-2105. URL http://www.ncbi.nlm.nih.gov/pubmed/18983678.

B. T. SHERMAN, D. W. HUANG, Q. TAN, Y. GUO, S. BOUR, D. LIU, R. STEPHENS, M. W. BASELER, H. C. LANE et R. A. LEMPICKI : David knowledgebase : a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics, 8:426, 2007. ISSN 1471-2105. URL http://www.ncbi.nlm.nih.gov/pubmed/17980028.

S. SHIBATA, M. SASAKI, T. MIKI, A. SHIMAMOTO, Y. FURUICHI, J. KATAHIRA et Y. YONEDA : Exportin-5 orthologues are functionally divergent among species. Nucleic Acids Res, 34 (17):4711–21, 2006. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/16963774.

D. SOH, D. DONG, Y. GUO et L. WONG : Consistency, comprehensiveness, and compatibility of pathway databases. BMC Bioinformatics, 11:449, 2010. ISSN 1471-2105. URL http://www.ncbi.nlm.nih.gov/pubmed/20819233. (pages 30, 110 et 137)

E. SOURY, E. OLIVIER, D. SIMON, P. RUMINY, K. KITADA, M. HIRON, M. DAVEAU, Y. BOYD, T. SERIKAWA, J. L. GUENET et J. P. SALIER : Chromosomal assignments of mammalian genes with an acute inflammation-regulated expression in liver. Immunogenetics, 53 (8):634–42, oct. 2001. ISSN 0093-7711. URL http://www.ncbi.nlm.nih.gov/pubmed/11797096.

J. STAMLER, D. WENTWORTH et J. D. NEATON : Is relationship between serum cholesterol and risk of premature death from coronary heart disease continuous and graded ? findings in 356,222 primary screenees of the multiple risk factor intervention trial (mrfit). JAMA, 256(20):2823–8, nov. 1986. ISSN 0098-7484. URL http://www.ncbi.nlm.nih.gov/pubmed/3773199. (page 23)

R. STEVENS, M. EGAÑA ARANGUREN, K. WOLSTENCROFT, U. SATTLER, N. DRUMMOND, M. HORRIDGE et A. RECTOR : Using OWL to model biological knowledge. International Journal of Human Computer Studies, 65(7):583–594, 2007.

M. D. STOBBE, G. A. JANSEN, P. D. MOERLAND et A. H. C. van KAMPEN : Knowledge representation in metabolic pathway databases. Brief Bioinform, nov. 2012. ISSN 1477-4054. URL http://www.ncbi.nlm.nih.gov/pubmed/23202525.

A. SUBRAMANIAN, P. TAMAYO, V. K. MOOTHA, S. MUKHERJEE, B. L. EBERT, M. A. GILLETTE, A. PAULOVICH, S. L. POMEROY, T. R. GOLUB, E. S. LANDER et J. P. MESIROV : Gene set enrichment analysis : a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 102(43):15545–50, oct. 2005. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pubmed/16199517.

Z. TENG, M. GUO, X. LIU, Q. DAI, C. WANG et P. XUAN : Measuring gene functional similarity based on group-wise comparison of go terms. Bioinformatics, 29(11):1424–32, juin 2013. ISSN 1367-4811. URL http://www.ncbi.nlm.nih.gov/pubmed/23572412.

THE GENE ONTOLOGY CONSORTIUM : Gene ontology : tool for the unification of biology. Nature genetics, 25(1):25–29, 2000.

THE GENE ONTOLOGY CONSORTIUM : Gene ontology annotations and resources. Nucleic acids research, 41(D1):D530–D535, 2012.

P. D. THOMAS, V. WOOD, C. J. MUNGALL, S. E. LEWIS, J. A. BLAKE et GENE ONTOLOGY CONSORTIUM : On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs : A short report. PLoS Comput Biol, 8(2):e1002386, 2012. ISSN 1553-7358. URL http://www.ncbi.nlm.nih.gov/pubmed/22359495.

Y. TSURUOKA, M. MIWA, K. HAMAMOTO, J. TSUJII et S. ANANIADOU : Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics, 27(13): i111–9, juil. 2011. ISSN 1367-4811.

Y. VAN DE PEER, J. S. TAYLOR, J. JOSEPH et A. MEYER : Wanda : a database of duplicated fish genes. Nucleic Acids Res, 30(1):109–12, jan. 2002. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/11752268.

B. C. VANTERU, J. S. SHAIK et M. YEASIN : Semantically linking and browsing pubmed abstracts with gene ontology. BMC Genomics, 9(Suppl 1):S10, 2008.

K. VERSPOOR, D. DVORKIN, K. B. COHEN et L. HUNTER : Ontology quality assurance through analysis of term transformations. Bioinformatics (Oxford, England), 25(12):i77–i84, 2009.

A. J. VILELLA, J. SEVERIN, A. URETA-VIDAL, L. HENG, R. DURBIN et E. BIRNEY : Ensemblcompara genetrees : Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res, 19(2):327–35, fév. 2009. ISSN 1088-9051. URL http://www.ncbi.nlm.nih.gov/pubmed/19029536.

J. H. VOGEL, A. von HEYDEBRECK, A. PURMANN et S. SPERLING : Chromosomal clustering of a human transcriptome reveals regulatory background. BMC Bioinformatics, 6:230, 2005. ISSN 1471-2105. URL http://www.ncbi.nlm.nih.gov/pubmed/16171528.

J. Z. WANG, Z. DU, R. PAYATTAKOOL, P. S. YU et C.-F. CHEN : A new method to measure the semantic similarity of go terms. Bioinformatics, 23(10):1274–81, mai 2007. ISSN 1367-4811. URL http://www.ncbi.nlm.nih.gov/pubmed/17344234. (pages 44 et 46)

L. WANG, P. JIA, R. D. WOLFINGER, X. CHEN et Z. ZHAO : Gene set analysis of genome-wide association studies : methodological issues and perspectives. Genomics, 98 (1):1–8, juil. 2011. ISSN 1089-8646. URL http://www.ncbi.nlm.nih.gov/pubmed/21565265.

P. L. WHETZEL, N. F. NOY, N. H. SHAH, P. R. ALEXANDER, C. NYULAS, T. TUDORACHE et M. A. MUSEN : Bioportal : enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic Acids Res, 39(Web Server issue):W541–5, juil. 2011. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/21672956. (page 34)

E. J. B. WILLIAMS et D. J. BOWLES : Coexpression of neighboring genes in the genome of arabidopsis thaliana. Genome Res, 14(6):1060–7, juin 2004. ISSN 1088-9051. URL http://www.ncbi.nlm.nih.gov/pubmed/15173112.

K. WOLSTENCROFT, P. LORD, L. TABERNERO, A. BRASS et R. STEVENS : Protein classification using ontology classification. Bioinformatics, 22(14):e530–8, juil. 2006. ISSN 1367-4811. URL http://www.ncbi.nlm.nih.gov/pubmed/16873517. (page 34)

X. WU, E. PANG, K. LIN et Z.-M. PEI : Improving the measurement of semantic similarity between gene ontology terms and gene products : insights from an edge- and ic-based hybrid method. PLoS One, 8(5):e66745, 2013. ISSN 1932-6203. URL http://www.ncbi.nlm.nih.gov/pubmed/23741529.

Z. WU et M. PALMER : Verb semantics and lexical selection. In Proc. of the 32nd annual meeting on Association for Computational Linguistics, p. 133–138, 1994. (page 44)

H. XU, F. LIU, N. SRAKAEW, C. KOPPISETTY, P.-G. NYHOLM, E. CARMONA et N. TANPHAICHITR : Sperm arylsulfatase a binds to mzp2 and mzp3 glycoproteins in a nonenzymatic manner. Reproduction, 144(2):209–19, août 2012. ISSN 1741-7899. URL http://www.ncbi.nlm.nih.gov/pubmed/22685254. (page 135)

M. XU et P. R. COOK : The role of specialized transcription factories in chromosome pairing. Biochim Biophys Acta, 1783(11):2155–60, nov. 2008a. ISSN 0006-3002. URL http://www.ncbi.nlm.nih.gov/pubmed/18706455.

M. XU et P. R. COOK : Similar active genes cluster in specialized transcription factories. J Cell Biol, 181(4):615–23, mai 2008b. ISSN 1540-8140. URL http://www.ncbi.nlm.nih.gov/pubmed/18490511.

Y. YAMAMOTO et T. TAKAGI : Biomedical knowledge navigation by literature clustering. J Biomed Inform, 40(2):114–30, avr. 2007. ISSN 1532-0480.

M. YANDELL et D. ENCE : A beginner's guide to eukaryotic genome annotation. Nat Rev Genet, 13(5):329–42, mai 2012. ISSN 1471-0064. URL http://www.ncbi.nlm.nih.gov/pubmed/22510764.

H. YANG, T. NEPUSZ et A. PACCANARO : Improving go semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. Bioinformatics (Oxford, England), 28(10):1383–1389, 2012. In press.

L. YAO, A. DIVOLI, I. MAYZUS, J. A. EVANS et A. RZHETSKY : Benchmarking ontologies : bigger or better ? PLoS computational biology, 7(1):e1001055, 2011.

S. YOO et J. CHOI : On the query reformulation technique for effective medline document retrieval. J Biomed Inform, 43(5):686–93, oct. 2010. ISSN 1532-0480.

G. YU, F. LI, Y. QIN, X. BO, Y. WU et S. WANG : Gosemsim : an r package for measuring semantic similarity among go terms and gene products. Bioinformatics, 26(7):976–8, avr. 2010. ISSN 1367-4811. URL http://www.ncbi.nlm.nih.gov/pubmed/20179076.

H. YU, L. GAO, K. TU et Z. GUO : Broadly predicting specific gene functions with expression similarity and taxonomy similarity. Gene, 352:75–81, 2005. (page 45)

B. ZHANG, D. SCHMOYER, S. KIROV et J. SNODDY : Gotree machine (gotm) : a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies. BMC Bioinformatics, 5:16, fév. 2004a. ISSN 1471-2105. URL http://www.ncbi.nlm.nih.gov/pubmed/14975175.

H. ZHANG, K.-H. PAN et S. N. COHEN : Senescence-specific gene expression fingerprints reveal cell-type-dependent physical clustering of up-regulated chromosomal loci. Proc Natl Acad Sci U S A, 100(6):3251–6, mars 2003a. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pubmed/12626749.

J. ZHANG : Evolution by gene duplication : an update. Trends in ecology & evolution, 18 (6):292–298, 2003.

P. ZHANG, Z. GU et W.-H. LI : Different evolutionary patterns between young duplicate genes in the human genome. Genome Biol, 4(9):R56, 2003b. ISSN 1465-6914. URL http://www.ncbi.nlm.nih.gov/pubmed/12952535.

Z. ZHANG, J. GU et X. GU : How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution ? Trends Genet, 20(9):403–7, sept. 2004b. ISSN 0168-9525. URL http://www.ncbi.nlm.nih.gov/pubmed/15313547.

Y. ZHAO, J. DONG et T. PENG : Ontology classification for semantic-web-based software engineering. IEEE T. Services Computing, 2(4):303–317, 2009. (page 34)

Q. ZHENG et X.-J. WANG : Goeast : a web-based software toolkit for gene ontology enrichment analysis. Nucleic Acids Res, 36(Web Server issue):W358–63, juil. 2008. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/18487275.