Prise en compte des connaissances du domaine dans l'analyse transcriptomique : Similarité sémantique, classification fonctionnelle et profils flous : application au cancer colorectal

Sid-Ahmed Benabderrahmane

▶ **To cite this version:**

Sid-Ahmed Benabderrahmane. Prise en compte des connaissances du domaine dans l'analyse transcriptomique : Similarité sémantique, classification fonctionnelle et profils flous : application au cancer colorectal. Informatique [cs]. Université Henri Poincaré - Nancy 1, 2011. Français. NNT : 2011NAN10097 . tel-01746248

HAL Id: tel-01746248
https://hal.univ-lorraine.fr/tel-01746248

Submitted on 29 Mar 2018

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

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

AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

Contact : ddoc-theses-contact@univ-lorraine.fr

LIENS

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10
http://www.cfcopies.com/V2/leg/leg_droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

Département de formation doctorale en informatique **École doctorale IAEM Lorraine**
UFR Sciences & Technologies

Prise en compte des connaissances du domaine dans l'analyse transcriptomique : Similarité sémantique, classification fonctionnelle et profils flous.

Application au cancer colorectal.

THÈSE

présentée et soutenue publiquement le 15 Décembre 2011

pour l'obtention du

Doctorat de l'université Henri Poincaré – Nancy 1

(spécialité informatique)

par

Sidahmed Benabderrahmane

Composition du jury

Président :	Amédéo Napoli	Directeur de recherche, CNRS, Loria, Nancy.
Rapporteurs :	Younès Bennani	Professeur, LIPN Université Paris-Nord.
	Christine Froidevaux	Professeur, LRI Université Paris-Sud.
Examinateurs :	Charles Auffray	Directeur de recherche, CNRS, Paris.
	Maude Pupin	Maitre de conférence, LIFL Université de Lille.
	Samy Tindel	Professeur, U. Henri Poincaré IECN, Nancy.
	Olivier Poch	Directeur de recherche, CNRS, Strasbourg.
	Marie-Dominique Devignes	Chargée de recherche, CNRS, Loria, Nancy.
Invité(e) :	Malika Smail-Tabbone	Maitre de conférence, U. Henri Poincaré, Loria, Nancy.

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thloria.

Remerciements

Je voudrais tout d'abord remercier M. Younès Bennani et Mme Christine Froidevaux pour leur disponibilité et d'avoir accepté de juger ce manuscrit.
Je remercie aussi M. Charles Auffray, M. Samy Tindel, Mme Maude Pupin d'avoir accepté de participer au jury de cette thèse.
Je tiens à remercier également M. Amedeo Napoli qui m'a donné l'occasion de faire partie de son équipe d'Orpailleurs, de m'avoir soutenu dans les moments difficiles, Grazie mille Amedeo.
Je remercie aussi M. Olivier Poch mon co-directeur de thèse, Wolfgang Raffelsberger et Hoan Ngoc Nguyen de leurs précieux échanges et enseignements en bioinformatique.
Je n'oublierai jamais M. Antoine Tabbone, qui m'a transféré un certain mail. Si je suis en train d'écrire cette page dans ce manuscrit, c'est grâce à ton click Antoine.
Je suis trés reconnaissant à l'INCa et la ligue contre le cancer de m'avoir accordé la bourse de thèse et la convention complémentaire qui m'a permis matériellement de poursuivre ma formation et de réaliser tous ces travaux. Merci aussi à Sébastien Retter pour son aide dans les justificatifs et les contrats.
J'en profite pour remercier aussi toutes les personnes qui travaillent dans le service des étrangers à la préfecture de Meurthe et Moselle, à l'Office d'immigration (OFII) de Metz et de Nancy, ainsi qu'à l'inspection de travail de Vandoeuvre.
Je tiens à ne pas oublier tous ceux qui ont partagé leur quotidien avec moi au Loria, j'ai passé des moments agréables avec vous. Je pense à Nizar, Birama, Emmanuel, Renaud (allez le PSG, staghfir-allah, allez l'OM), Jean-François (merci pour les conseils info), Houari, Yahia, Elias, Chedy (people of Gaza we are with you), Mohamed, Yakoub, Tarek, Bilel, Sofiane, Fares, Hamza, Yacine, Adam, Ibrahim,...
Je remercie aussi tous mes amis d'enfance en Algérie, Abdelmouneim, Jamal, Ali, Boucif, Dino, Abdelaziz, Abdallah, Abdli, Boumedien, Said, Djillali, Boubakar et tous les autres amis. Rendez-vous, l'été prochain incha'allah, à la plage Hafer-Jmel.
Je remercie tous les membres des familles Benabderrahmane, Mekami et Slimani. Merci pour votre soutien. Merci mon père pour ton encouragement. Merci ma mère, je n'oublierai jamais quand tu gâchais ton sommeil pour me réveiller. Vous restez dans mon coeur.
Merci à chaque personne qui m'a enseigné une lettre, de l'école primaire jusqu'à maintenant.
Last but not least, Malika et Marie-Dominique, je n'arrive pas à trouver les mots justes pour vous remercier de votre soutien, de votre accompagnement, de vos conseils, de vos enseignements tant scientifiques qu'humains. Je garderai cela en mémoire. Thanmirt Malika, et merci du fond du coeur Marie-Dominique.

Je dédie cette thèse
À mes Parents,
À ma ZZZzzzzzouzou ;-),
À mon Frère Souleymane,
À mes Soeurs,
À mes Grands Parents,
À toute ma famille,
À mes amis,
À toi : Sarah ou Maria la sainte (si tu es une fille),
Mohamed ou Isaac (si tu es un garçon).
,
.
.
....
.
.
À la Mémoire de...
Amirouche Aït Hamouda,
Abdelhamid Ibn Badis.
.
.
.
.
"Lorsque nous serons libres, il se passera des choses terribles. On oubliera toutes les souffrances de notre peuple, pour se disputer les places. Nous sommes en pleine guerre et certains y pensent déjà. Oui j'aimerais mourir au combat avant la fin." (Larbi Ben M'hedi 1923-1957)

Table des matières

Table des figures

Liste des tableaux

Introduction générale

1 Contexte biologique et applicatif :

Cette thèse s'inscrit dans le cadre d'une collaboration entre le laboratoire de bioinformatique et de génomique intégrative (LBGI) de l'IGBMC à Strasbourg, et l'équipe Orpailleur du LORIA à Nancy, en partenariat avec l'INCa (Institut National du Cancer). Le sujet de thèse pose le problème de la fouille de données complexes en tenant compte des connaissances a priori du domaine. Les données considérées sont des résultats d'analyses transcriptomiques pour le cancer colo-rectal.

2 Problématique :

Les techniques des puces à ADN permettent de mesurer le niveau d'expression de plusieurs milliers de gènes dans diverses situations. Le niveau d'expression d'un gène reflète son activité et préfigure la quantité de protéines codées par ce gène qui sera produite dans la cellule. Des variations de niveau d'expression peuvent être observées pour un même gène en fonction de l'état physiologique du tissu étudié. Pour des échantillons prélevés sur des sujets malades, l'objectif de l'analyse des niveaux d'expression des gènes est de repérer les gènes qui présentent des variations significatives par rapport aux prélèvements sur sujet sain. De nombreuses méthodes ont été développées pour faciliter cette analyse. Les difficultés qui demeurent concernent l'interprétation des résultats d'analyse, c'est-à-dire la compréhension des raisons pour lesquelles l'expression de ces gènes est modifiée et des mécanismes moléculaires de ces dérèglements. C'est pourtant cette étape qui peut ouvrir la voie à la conception de nouvelles stratégies de diagnostic ou de thérapie.

3 Objectif général et contributions :

L'interprétation de données transcriptomiques repose sur deux types d'analyse : la construction de profils d'expression des gènes d'une part, et d'autre part l'étude des annotations fonctionnelles, généralement issues de l'ontologie GO (Gene Ontology) des gènes. Le deuxième type d'analyse sert à donner un sens au premier. J'ai voulu contribuer à l'efficacité de ce processus d'interprétation à quatre niveaux.

– L'utilisation d'une mesure de similarité pour la classification fonctionnelle de gènes signatures n'était pas effective jusqu'à présent car la plupart des mesures de similarité sémantique n'étaient pas bien adaptées, en particulier elles ne respectaient pas la propriété de maximalité selon laquelle la similarité d'un objet avec lui même est maximale (égale à 1

pour une mesure normée). J'ai donc défini une nouvelle mesure de similarité sémantique et fonctionnelle (IntelliGO), robuste et capable de considérer différemment les annotations fonctionnelles selon les différents aspects de GO, selon les espèces et aussi les codes d'évidence. Cette partie de travail a fait l'objet d'une publication dans une revue internationale [BSTP$^+$10].

– La validation de la classification fonctionnelle grâce à des jeux de données constitués d'ensembles de référence m'a conduit à développer une méthode d'analyse de recouvrement entre clusters fonctionnels et ensembles de références. Cette contribution a fait l'objet de deux communications orales dans [BSTND, BSTP$^+$ce].
– Cette méthode d'analyse de recouvrement peut alors être appliquée au recouvrement entre clusters fonctionnels et profils d'expression, notamment dans le cas d'un jeu de données relatif au cancer colo-rectal. Pour cela la notion de profil d'expression est revisitée avec l'introduction des profils d'expression différentielle flous. Ce travail a été présenté dans [BDST$^+$09].
– La mesure de similarité sémantique (IntelliGO) est ensuite généralisée à tout vocabulaire structuré en rDAG. La possibilité de réaliser alors un clustering sémantique des termes de l'ontologie a été utilisée sur l'ontologie MedDRA en vue de faciliter la fouille de données biomédicales. Cette partie a été présentée dans [BBST$^+$11].

Etat de l'art

Chapitre 1

État de l'art : analyse transcriptomique et connaissances du domaine

Sommaire

L'objectif de ce chapitre est la présentation des données issues des expériences de puces à ADN ainsi que l'introduction aux connaissances du domaine pouvant guider l'analyse de ces données. Cette technologie a pour finalité la compréhension des mécanismes de régulation de l'expression de gènes dans un génome. A cet effet, les biologistes réalisent des expérimentations sur différents types d'échantillons pour identifier les gènes qui s'expriment de façon différentielle, réagissant ainsi aux différents stimuli externes. Le cas idéal serait de pouvoir regrouper les gènes ayant le même comportement du niveau d'expression afin de pouvoir prédire les catégories fonctionnelles et les mécanismes biologiques qui sont associés aux gènes. Eventuellement, des approches de data mining seront appliquées afin d'extraire des connaissances à partir de ce type de données.
Dans ce chapitre seront présentées de manière non exhaustive la technique de puces à ADN permettant de mesurer le niveau d'expression des gènes, les approches statistiques de mise en évidence de gènes différentiellement exprimés, la classification des gènes en profils d'expression, les annotations fonctionnelles de gènes et l'analyse fonctionnelle des données d'expression.

Mots clés : puces à ADN, niveau d'expression, profil d'expression de gènes, classification de gènes, annotation fonctionnelle de gènes, analyse fonctionnelle.

1 Du génome au transcriptome : les grandes étapes de la génomique fonctionnelle

Dans le monde du vivant, tous les organismes à l'exception des virus sont composés de cellules qui représentent l'unité élémentaire de reproduction d'un être vivant. Les cellules comportent un ou plusieurs compartiments délimités par une membrane composée de lipides, et un génome composé d'acides nucléiques renfermant des instructions de structuration et de reproduction de l'organisme vivant [Kni06].
L'ensemble de ces instructions est ordonné de façon linéaire sur les différents chromosomes de chaque cellule, où chaque gène du génome occupe une place qui s'appelle le locus. Depuis les expériences fondatrices de MacLeod, McCarthy, et d'Avery [AMM44], on sait que ce matériel génétique est le plus souvent (sauf chez certains virus) l'Acide DésoxyriboNucléique (ADN), qui est le support de l'information génétique. Sa structure est universelle, seule la longueur de la molécule et l'enchainement de ses constituants varient selon les espèces.
L'ADN est structuré en une hélice à deux brins se faisant face, dont chacun est composé d'une chaîne de désoxyribonucléotides (bases) abrégés en A, T, C et G pour la Thymine, la Cytosine, l'Adénine et la Guanine respectivement (Figure 1). Chacun des brin de l'ADN contient la séquence d'un gène tandis que l'autre brin contient la séquence complémentaire. Ceci est possible car les nucléotides trouvés dans un brin possèdent des nucléotides complémentaires avec lesquels ils peuvent interagir par des liaisons hydrogènes, suivant les couples possibles : A-T et T-A, dans le cas de la thymine qui est toujours en face d'une adénine ; C-G et G-C, dans le cas d'une cytosine qui est toujours en face d'une guanine. Par exemple la molécule d'ADN humaine est gigantesque et contient 3,2 milliards de paires des quatre bases réparties sur 23 paires de chromosomes. L'ordre d'apparence des bases détermine en grande partie les caractéristiques physiques des être vivants et renferme l'information nécessaire à leur développement et à leur survie [WC53, GBee06]. De plus,l'environnement et la nutrition jouent un rôle non négligeable, appelé "épigénétique", sur la façon dont cette information s'exprime [ea06].
Les gènes d'une cellule vivante interagissent entre eux et réagissent aux événements externes (nutriments, médicaments, hormones, ...) par des mécanismes d'activation ou de répression. On dit que le gène "s'exprime" pour donner naissance à un phénotype, c'est à dire une protéine dotée d'un certain nombre d'Acide RiboNucléique (Figure 2). En effet, quand un gène est activé par un événement externe (i.e. lorsqu'il s'exprime), il se recopie par le biais d'un brin d'ARN (Acide RiboNucléique) par un processus chimique appelé transcription, et cette copie est appelée ARN messager ou ARNm [LY00, Wei01]. La molécule d'ARNm joue un rôle primordial en tant que messager portant l'information du noyau d'une cellule vers le cytoplasme qui est à terme le lieu où sont synthétisées les protéines. Le procédé de synthèse de protéine s'appelle la "traduction", et se fait au sein de structures complexes appelées ribosomes, dont le rôle principal est de synthétiser les protéines en décodant l'information contenue dans l'ARN messager. On peut considérer donc la molécule d'ARNm comme une molécule du vivant, située entre deux mondes : un génome passif et identique pour toutes les cellules, et un protéome en changement perpétuel [JFB04].
Les protéines sont définies comme étant une succession linéaire d'acides aminés. La séquence en base de l'ARN détermine la séquence d'acides aminés présents dans la protéine. La relation entre la séquence des nucléotides d'un gène et la séquence des acides aminés dans la protéine

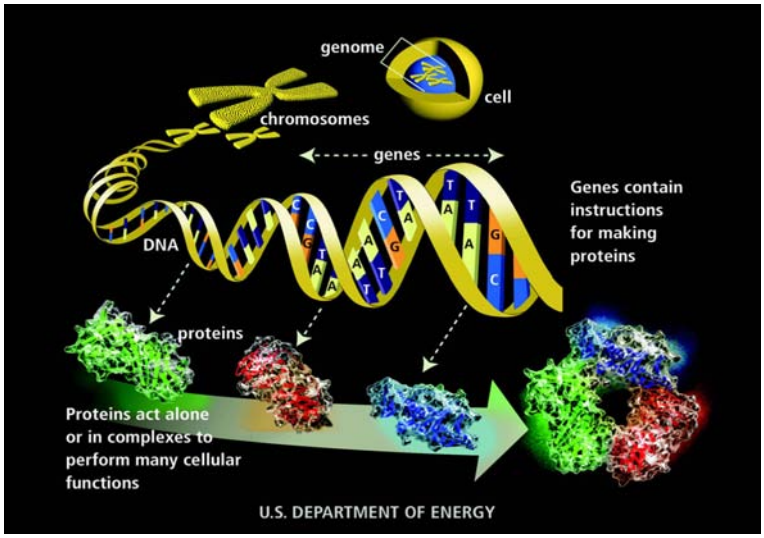


FIGURE 1 – Illustration d'une cellule d'un être vivant, dont le génome serait composé d'un chromosome. Ce chromosome, contient une molécule d'ADN compactée. Un gène représente une partie de la molécule d'ADN du chromosome. Un ADN est représenté par les deux brins formant une double hélice. Il se compose d'une suite de nucléotides. Chaque nucléotide est dénommé par l'initiale du nom de la base azotée qui le compose : A (Adénine), C (Cytosine), G (Guanine) et T (Thymine). Les gènes contiennent les instructions nécessaires à la fabrication des protéines.

correspondante est donnée par le code génétique. En effet, chaque acide aminé est codé par des triplets de nucléotides, dont certains triplets codent pour le même acide aminé. Une réaction de biosynthèse de protéines consiste à faire glisser le ribosome le long du brin d'ARN messager afin de lire la succession de triplets, pour construire la chaîne d'acides aminés correspondante, jusqu'à synthèse totale de la protéine. Le résultat de la réaction est alors une nouvelle chaîne d'acides aminés, qui quitte ensuite le ribosome et se replie sur elle-même dans une configuration caractéristique, qui est déterminée par la séquence des acides aminés. C'est la forme tridimensionnelle de la protéine qui détermine son activité c'est à dire sa fonction chimique à l'intérieur de l'organisme.

Il parait donc que les gènes contrôlent le processus de formation des protéines, composés fondamentaux pour tous les processus biologiques par le biais des molécules d'ARNm. D'ailleurs, les protéines ne sont pas seulement les composants majeurs de la plupart des structures cellulaires, mais elles contrôlent aussi presque la totalité des réactions chimiques qui existent dans l'organisme du vivant.

Toutes les cellules présentes dans un organisme vivant contiennent pratiquement les mêmes gènes, mais la différence réside dans le fait qu'elles synthétisent des protéines différentes. Ceci implique que ces cellules devraient avoir une combinaison particulièrement différente de gène actifs, c'est à dire que quand certains d'entre eux sont dans un état actif (s'expriment), d'autres restent inactifs. Par exemple, pour le génome humain qui contient approximativement 30 000 gènes produisant environ 200 000 ARNm, c'est seulement quelques milliers de ces gènes qui vont s'exprimer pour produire une protéine dans une cellule donnée dans le corps humain [Kni06, JFB04].

Depuis deux décennies, et avec l'arrivée des techniques de séquençage des génomes [Ven01], la génomique fonctionnelle a vu le jour en permettant l'étude à grande échelle des fonctions biologiques de gènes de différents organismes [Pev09]. Dans ce cadre, l'étude plus particulièrement de l'ARN transcrit à partir des gènes exprimés du génome est appelée l'étude du transcriptome.

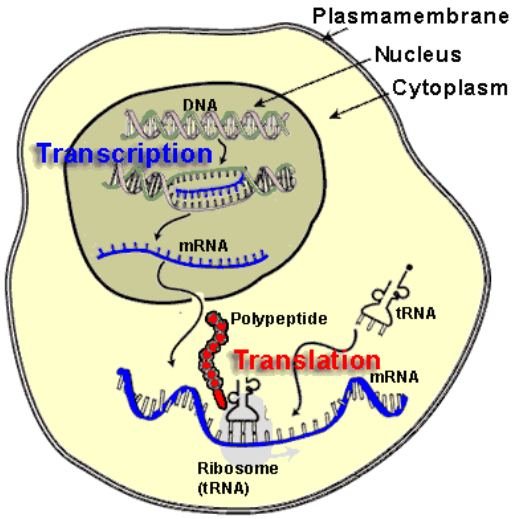


FIGURE 2 – Dogme central de la biologie moléculaire. L'information génétique qui constitue le génotype d'un organisme s'exprime pour donner naissance à un phénotype, c'est à dire l'ensemble des caractères de cet organisme. Cette expression du génome se fait en interaction avec divers facteurs de l'environnement (nutriments, lumière...). Elle se fait en plusieurs étapes : 1. La transcription, qui est le transfert de l'information génétique de l'ADN vers une autre molécule, l'ARN. 2. La traduction (en anglais translation), qui est un transfert d'information depuis l'ARN vers les protéines. 3. L'activité des protéines. L'activité des protéines détermine l'activité des cellules, qui vont ensuite déterminer le fonctionnement des organes et de l'organisme.

Il faut savoir que le changement d'état d'une cellule se fait par l'expression des gènes qui s'y trouvent, et que ces derniers n'ont pas forcément tous le même niveau d'expression. Pour cela on parle souvent des différences des niveau d'expression au sein d'une même cellule [Cla99, JEB+97]. En effet, si on prend l'exemple de cellules cancéreuses, la division cellulaire est en état accéléré, et les gènes cibles produisent beaucoup d'ARNm afin de synthétiser les protéines. Ainsi, la quantité d'ARNm produite par un gène présent dans un tissu sain doit être différente de celle produite par le même gène dans le tissu tumoral. C'est pour cela que les biologistes s'intéressent aux gènes qui sont différentiellement exprimés dans un tissu par rapport à un autre, dans des tissus provenant de tumeurs ou non [BSE+04, BFR+04, RJGS05, KKR+04].

2 L'analyse transcriptomique

2.1 Les dispositifs expérimentaux

Selon le type de tissu ou de cellule (musculaire, neurone,...) et selon l'état physiologique (sain, cancéreux,...) les gènes s'expriment en donnant des molécules ARNm différentes et avec des quantités variables. La tâche d'identification de la quantité d'ARNm produite devient donc une tâche très difficile, qu'il faudra répéter pour de multiples échantillons. Des technologies très sophistiquées ont été développées pour améliorer la sensibilité et la reproductibilité nécessaire afin de limiter au maximum les erreurs de mesures dans ce monde minuscule. Ces technologies peuvent se répartir en deux groupes selon qu'elles sont fondées sur le séquençage ou sur l'hybridation.

2.1.1 Techniques fondées sur le séquençage

La première approche est basée sur le séquençage de portions de séquences de l'*Acide Désoxyribo Nucléique Complémentaire* (ADNc) d'une cellule particulière. L'ADNc d'une cellule ou d'un tissu particulier est produit par transcription inverse (*reverse transcription*) [BWPS78, Bus00] de l'ARNm extrait de cette cellule ou de ce tissu. Cette copie d'ADN est plus aisément séquençable que l'ARN lui même. Les sous-séquences obtenues par séquençage à grande échelle sont appelées *Expressed Sequence Tag* : EST, pour étiquettes de séquences exprimées, et elles peuvent être utilisées pour quantifier les ARNm transcrits [AKG+91, WLB10]. Une amélioration de la méthode EST a été proposée afin de remédier au problème de taux d'erreur élevé lors du séquençage. Cette technologie améliorée s'appelle l'analyse sérielle de l'expression génique (SAGE) [VZVK95]. Dans ce cas, les molécules d'ADNc sont manipulées pour produire une molécule d'ADN synthétique comportant des étiquettes de tous ces ADNc. Le séquençage de cette molécule synthétique produit des quantités de séquences proportionnelles aux quantités d'ARNm de départ. Plus récemment, les méthodes de séquençage de l'ADN ultra-performantes, dites "de nouvelle génération", ont permis le développement des techniques RNAseq, conduisant à l'accumulation rapide et massive de données de séquence reflétant l'expression des gènes [CG08].

2.1.2 Techniques fondées sur l'hybridation

La deuxième approche de quantification du niveau d'expression des gènes est basée sur une technique d'hybridation complémentaire de deux brins des molécules d'ADN [Hel02, Yan09].Depuis les années 90, les technologies des puces à ADN (dites à haut débit) se sont développées, offrant au fil des années des capacités et des précisions croissantes pour mesurer la distribution des transcrits dans une cellule. On parle de "mesure du transcriptome" [Sto05, JFB04].
Le principe fondamental de cette technique repose sur l'utilisation de molécules d'ADN déposée sur plusieurs milliers d'unités d'hybridation à la surface d'un support de quelques cm^2 (lame de microscope), comme illustré dans la Figure 3. Chacune de ces unités d'hybridation est composée de plusieurs milliers d'exemplaires d'une même molécule d'ADN figée sur le support solide de la puce. La première étape pour mesurer le niveau d'expression d'un gène d'une cellule donnée consiste à générer à partir des molécules ARNm de cette cellule, des ADNc par la méthode de transcription inverse. Ces ADNc sont ensuite couplés avec des marqueurs fluorescents (Cy5, Cy3) ou radioactifs. Dans le cas où une complémentarité existe entre une sonde donnée (fragment d'ADN sur la puce) et un ADNc cible marqué, alors un couple ADNc/sonde est formé par hybridation (appariement).
Après un certain laps de temps, le moment où les réactions chimiques d'hybridation atteignent un stade final, des spots fluorescents apparaissent sur le support de la puce permettant ainsi d'identifier et de quantifier avec un scanner par laser, sur chaque sonde, l'intensité du signal émis (radioactivité ou fluorescence). L'intensité du signal d'hybridation sur chaque sonde, reflète par hypothèse la quantité relative du transcrit produite par chaque gène dont la séquence correspond à cette sonde. Ainsi, une sonde qui émet un fort signal d'hybridation correspond à un gène fortement exprimé. Des traitement de normalisation des signaux d'hybridation sont nécessaires pour pouvoir interpréter ces mesures [PPCP07, PCGP06]. La technologie des puces à ADN permet donc de récupérer les niveaux d'expression mesurés pour chaque gène dans un certain nombre de situations. Ces valeurs d'expression constituent ce que l'on appelle classiquement le *profil d'expression* de ce gène pour un jeu de situations données. Ces situations peuvent être par exemple des tissus de différents types (sain, cancéreux,...), dans ce cas là on parle de

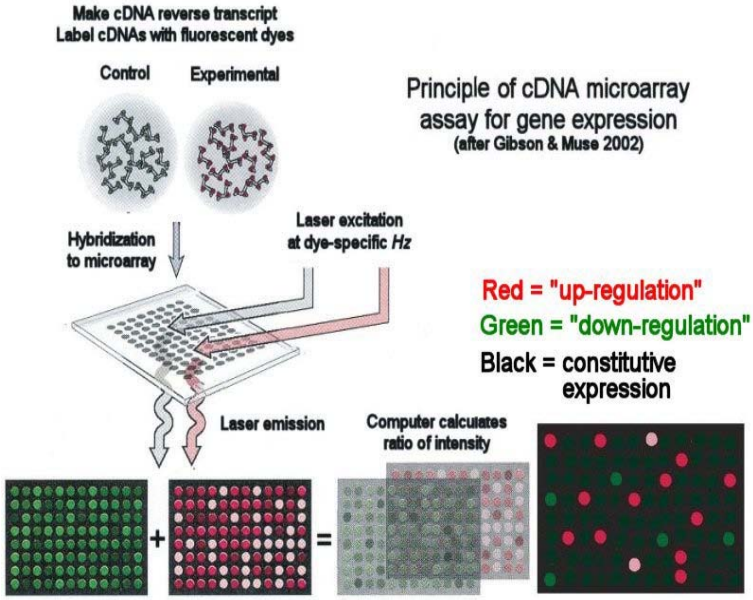


FIGURE 3 – Principe de fonctionnement des puces à ADN. Des fragments d'ADNc amplifiés par la technique de PCR ou des oligonucléotides sont déposés sur les sondes présentes sur une lame de microscope. Des ADNc sont générés à partir des échantillons étudiés, et marqués à l'aide de deux fluorochromes (Cy3 et Cy5). Le mélange est alors hybridé sur la puce. La puce est par la suite scannée par un laser à plusieurs longueurs d'onde, donnant ainsi une intensité numérique à chaque sonde. L'image produite est ensuite scannée et les données seront normalisées puis analysées et interprétées.

données d'expression statique [ESBB98]. Si ces situations sont des expériences produites dans le laboratoire à des instants différents, alors on parle de données d'expression dynamique [DDCG09].

2.2 Les grandes étapes de l'analyse de données d'expression

Les expériences avec les puces à ADN nous fournissent des données d'expression transcriptomique qui sont de nature numérique, et qui peuvent être annotées avec des annotations fonctionnelles (données symboliques). Les erreurs de mesures, liées à l'expérience, peuvent se répercuter sur la suite du processus d'analyse. Ainsi à l'instar de tout type d'approche expérimentale, ces données nécessitent des étapes de traitement statistique et de normalisation préalable, afin de réduire le bruit technique, pour pouvoir les exploiter de manière pertinente [TOR+01a, BL01]. Ces traitements permettent de restreindre les fluctuations aléatoires dues au artéfacts de mesures qui sont liés généralement au matériel utilisé ou bien à la variation biologique des tissus étudiés. Par exemple, selon le type de lame utilisée, des spots défectueux sont souvent observés après l'étape de l'acquisition des images par scanner. Pour cela, plusieurs produits de puces à ADN sont commercialisés dans le marché avec des solutions logicielles afin de supprimer des bruits engendrés pendant le processus de mesure des niveaux d'expression. Sur la plupart des puces, on y trouve aussi des réplicats de sondes, afin d'accéder à la variabilité locale du signal.
Globalement les grandes étapes de l'analyse de données d'expression, peuvent se résumer comme suit : Acquisition des données et filtrage, Traitements statistiques, Identification des gènes différentiellement exprimés, Extraction de profils d'expression et analyse fonctionnelle de gènes

[Mic08, GUE05, Meu05, IA05].

2.2.1 Étape 1 : Acquisition des données d'expression et filtrage

Une fois que l'hybridation a atteint son stade final, la lame contenant les sondes avec des granularités fluorescentes peut être analysée avec un scanner (Figure 3). Un algorithme de segmentation d'image est ensuite appliqué sur l'image résultante afin de localiser les pixels correspondants aux sondes et donner ensuite une valeur numérique représentant l'intensité du signal fluorescent.
Cette intensité du signal lumineux représente l'état primaire des données d'expression. Elle nécessite donc des traitements pour valider l'analyse en aval. Or, comme nous l'avons signalé, il se peut que certains spots soient défectueux. Pour cela une étape de filtrage est requise afin d'écarter les spots défectueux. Cette étape de filtrage considère à la fois la détection des spots non valides et leurs critères qualitatifs physiques et géométriques comme le niveau de saturation du signal ou le rapport signal sur bruit [Meu05, YBDS02].

2.2.2 Étape 2 : Traitements statistiques

Les expériences avec les puces à ADN sont généralement répétées plusieurs fois afin de tenir compte de la variabilité expérimentale. Un moyen physique pour la répétition des expériences est l'utilisation de plusieurs lames. Dans ce cas, une normalisation des données inter-lames (*scaling*) est nécessaire pour réduire la variance des mesures entre ces lames [MNS10, TOR$^+$01b, YBS01]. Cependant, on constate souvent après la répétition de ces expériences, l'apparition de valeurs aberrantes (outliers). Ces valeurs aberrantes ont pour cause généralement un problème physique lié à la qualité de la sonde dans la puce, ou un problème de bruit de fond à la surface des lames (erreur de segmentation,...). Plusieurs outils ont été proposés pour faire face à ce genre de problème, par exemple : DNA-Chip Analyzer (utilisé par nos collègues de l'IGBMC de Strasbourg), MADSCAN [1], MicroArray Data Suites of Computed ANalysis [Wu09, LMLB$^+$04, Dra03]. Bien que ces approches aient été proposées pour faire face au bruit de fond, ce dernier reste plus ou moins omniprésent à cause de plusieurs facteurs intrinsèques et extrinsèques, notamment ceux relatifs à la qualité de la technologie utilisée, et à l'efficacité des traitements statistiques appliqués. *Quackenbush et al.* confirment qu'il n'existe pas d'approche de normalisation universelle mais plutôt une solution pour chaque cas [Qua01].

2.2.3 Étape 3 : Mise en évidence de gènes différentiellement exprimés

L'un des principaux objectifs des expériences des puces à ADN est la mise en évidence de gènes différentiellement exprimés entre deux ou plusieurs types d'échantillons (situations ou conditions biologiques). L'intérêt majeur de l'étude de ces gènes est l'identification éventuelle de gènes pouvant servir de marqueurs de maladies ou de différenciation cellulaire.
Classiquement pour mesurer la différence d'expression d'un gène entre deux conditions (ex : tissu sain ou normal contre tissu cancéreux), on estime l'amplitude des variations (*fold change*) qui existent entre elles. Soit $Cy5_{normal}$ le niveau d'expression d'un gène dans un tissu normal et $Cy3_{cancer}$ son niveau d'expression dans le tissu cancéreux. On calcule l'amplitude de variation par la formule :

$$M_g = Log_2 \frac{Cy5_{normal}}{Cy3_{cancer}}. \quad (1)$$

1. MADSCAN, MicroArray Data Suites of Computed ANalysis (http ://www.madtools.org)

	Acceptation de $H_{0,g}$ (g négatif)	Rejet de $H_{0,g}$ (g positif)
$H_{0,g}$ vrai	1-α	α
$H_{1,g}$ vrai	β	1-β

TABLE 1 – Risques d'erreurs et degrés de confiance avec un test d'hypothèse nulle ($H_{0,g}$) ou non ($H_{1,g}$).

Pour une liste de gènes, on pose M_μ et M_σ pour la moyenne et l'écart type des ratios des niveaux d'expression calculés.
Cette valeur d'amplitude de variation est ensuite comparée à un seuil préfixé afin de déterminer si le gène s'exprime de façon différentielle ou pas, généralement un facteur de 2 (2-fold change) [Qua01]. Cependant, et comme nous l'avons signalé précédemment, le problème de bruit de fond est parfois difficilement évitable ce qui implique des erreurs que peuvent engendrer les méthodes de calcul de l'amplitude des variations d'expression. Un autre défi majeur qui intervient, c'est qu'il faudrait être capable de différencier les variations biologiques qui reflètent les processus biologiques de la cellule par rapport aux variations expérimentales [Mic08, GUE05]. A cet effet, des approches statistiques ont été proposées pour extraire les gènes différentiellement exprimés.
Tests paramétriques : Parmi les tests paramétriques utilisés pour identifier les gènes qui s'expriment de façon différentielle entre deux conditions biologique est le test de *Student* appelé aussi le *t-test* [TOTZ01]. Il s'agit d'un test classiquement utilisé en statistique pour évaluer la différence qui peut résider entre deux échantillons comparés. Ce test suppose que les données d'expression suivent une loi normale.
Ce test prend en considération des hypothèses posées sur le niveau d'expression des gènes étudiés. Il vérifie l'hypothèse nulle sur le jeu de données étudié. L'hypothèse nulle représente le cas où un gène g n'a pas de variation d'expression : $H_{0,g}$={La différence d'expression du gène g entre deux conditions est nulle}, avec une erreur α dite erreur de type 1 qui représente la probabilité d'obtenir un faux-positif (FP), c'est à dire un gène trouvé comme différentiellement exprimé alors qu'il ne l'est pas. Cette hypothèse est testée contre l'hypothèse alternative $H_{g,1}$={La différence d'expression du gène g entre deux conditions est non nulle}, qui à son tour est associée à une erreur β de type 2, reflétant la probabilité d'obtenir un faux-négatif (FN), c'est à dire un gène qui n'est pas trouvé comme différentiellement exprimé alors qu'il l'est. La Table 1 indique la probabilité de chaque situation possible. On appelle la valeur 1-α le *degré de confiance*, c'est la probabilité de ne pas obtenir de faux positifs. La quantité 1-β représente la *puissance* : c'est la probabilité de ne pas avoir un faux négatif. La valeur *p-value* représente la probabilité d'obtenir la valeur observée du ratio M_g des niveaux d'expression d'un gène g si l'hypothèse nulle est acceptée. Par conséquence plus cette valeur est petite moins la différence d'expression est due au hasard, et donc l'hypothèse nulle est rejetée. En effet, le gène est dit différentiellement exprimé si et seulement si cette *p-value*$\leq\alpha$, sinon le gène en question fait partie des faux positifs [Meu05].
Rappelons que les puces à ADN mesurent les niveaux d'expression de plusieurs milliers de gènes simultanément. Par conséquence, il faut faire autant de tests d'hypothèse nulles que de gènes présents dans le support de la puce. Pour une population de N gènes, la valeur de la *p-value* est ensuite estimée par le biais de la valeur T_g (2) qui est calculée et comparée à la distribution du test de *Student*.

$$T_g = \sqrt{N}\frac{M_\sigma}{M_\mu}. \tag{2}$$

Il parait donc clairement que plus le nombre de gènes analysés est grand plus le test est puissant. D'autres types de tests non paramétriques qui sont plus adaptés aux données d'expression bruités

ont été utilisés, par exemple le test SAM (Significance Analysis of Microarrays). Cette méthode se base sur un tirage aléatoire (booststrapping) du jeu de données initial et consiste à observer les données qui sont liées au hasard [Tib06, TTC01]. Une implémentation de ces approches est disponible dans la librairie *limma* du package *R bioconductor*[2].

L'approche bayésienne : Cette méthode a été utilisée pour la sélection de gènes différentiellement exprimés en se basant sur le théorème de *Bayes* [TTC01]. Le principe de base de ce théorème est l'estimation d'une probabilité a posteriori d'un évènement X en fonction de la probabilité a priori de l'évènement : $P(X|Y) = P(x) * \frac{P(Y|X)}{P(Y)}$.
Cette méthode a été proposée pour faire face aux problèmes rencontrés avec le test t, qui nécessite un nombre d'échantillons important [SS04, BL01].

Analyse de la variance (ANOVA) : Cette méthode est souvent utilisée dans des problèmes où le jeu de données étudié a de multiples facteurs (symptômes, âge, habitudes alimentaires, sexe,...). Il parait donc plus clairement que cette approche est adaptée aux puces à ADN, puisque celles-ci génèrent aussi des données avec de multiples facteurs de variabilité. L'approche ANOVA étudie l'impact associé à chaque facteur sur un échantillon par rapport à un jeu de données global qui suit une loi normale. Cette méthode est bien adaptée pour mesurer la différence d'expression d'un gène par rapport à plus de deux conditions biologiques [KMC00]. Au cours des expériences avec les puces à ADN, la variation d'expression pourrait être causée par plusieurs facteurs : variations des type de tissus étudiés, des individus, des fonctions des gènes. Pour cela cette méthode est utilisée pour modéliser chaque facteur en construisant un modèle statistique qui utilise la distribution a posteriori de chaque facteur [Pav03]. Une implémentation de la méthode est disponible dans la librairie *maanova* sur *R bioconductor*.
En résumé, les approches d'extraction de gènes différentiellement exprimés représentent la première étape nécessaire pour préparer ces données volumineuses en vue de leur appliquer des processus d'extraction de connaissance. Pour cette raison, cette étape requiert une bonne maîtrise des outils disponibles à cet effet, et de bonnes connaissances théoriques sur les tests statistiques pour pouvoir appliquer la méthode adéquate aux jeux de données dont on dispose.

2.2.4 Mise à disposition des données d'expression en vue de l'interprétation :

Après avoir préparé les données d'expression, il faut encore gérer ces gigantesques quantités de données hétérogènes et s'assurer d'un bon système de stockage et sauvegarde afin de les traiter par la suite. Entre autres, il faut assurer une traçabilité pendant toute la chaîne du traitement des images scannées, les informations relatives aux sujets cliniques (type de tissus), type de puces utilisées, condition d'hybridation, etc...
A cet effet, plusieurs bases de données existent, et qui permettent la gestion de ces données. Le format : MIAME (Minimum Information About a Microarray Experiment), est un data warehouse fourni par la fondation FGED (The Functional Genomics Data Society) [3] et qui spécifie un standard d'échange afin de rendre publiques des données d'expression pré-traitées et d'annoter avec le minimum d'informations requises pour le traitement [BA01]. Plusieurs bases de données suivent le format FGED (MAGE-ML : MicroArray Gene Expression-Markup Language) et fournissent aussi des données issues des puces à ADN : CIBEX (Center for Information Biology gene EXpression database) [IK03], The ArrayExpress Archive [ZBRPB10], Gene Expression Omnibus [BE06], MARS (microarray analysis, retrieval, and storage system) [MMS+05]. Citons également l'initiative française SIDR : "Standard Infrastructure for Distributed Data Repository",

2. http ://bioconductor.org
3. http ://www.mged.org/

	Condition 1	Condition 2	Condition 3	....	Condition p
Gene 1	314	445	90	1234	955
Gene 2	78	123	515	63	1700
Gene 3	.	.	.	.	.
...	.	.	.	.	.
Gene i	.	.	.	.	.
....	...	...	...	...	...
Gene n	400	612	19	46	150

TABLE 2 – Représentation standard des données transcriptomiques via une matrice d'expression (M). En ligne sont représentés les gènes étudiés dans la puce à ADN. Les colonnes représentent les conditions biologiques relatives à différentes expériences. Une cellule (i,j) représente le niveau d'expression du gène (i) dans la condition (j), le niveau d'expression étant le degré d'abondance de molécules ARNm produites par le gène dans la condition en question sur une échelle à définir.

qui s'adresse de façon plus générale à tout jeu de données haut-débit [4].

2.3 Étape 4 : Extraction et définition des profils d'expression des gènes

2.3.1 Définition des profils d'expression :

Une fois que les données d'expression sont extraites et filtrées, on est souvent confronté à une représentation standard qui est illustrée dans la Figure 2. Cette représentation est réalisée selon ce qu'on appelle une *matrice d'expression* ($M[n,p]$) de taille $n \times p$, et définie dans $\mathbf{R}_+$. Dans cette matrice sont représentés les (n) gènes étudiés dans la puce à ADN (lignes) qui s'expriment différentiellement dans les (p) conditions biologiques relatives à différentes expériences (colonnes). Une cellule $M[i,j]$ représente donc le niveau d'expression du gène (i) dans la condition (j).

Avec cette matrice d'expression, on pourrait appliquer les techniques de datamining pour extraire les connaissances sous-jacentes et aider le biologiste à avoir une meilleure interprétation de ces données. Le but étant de créer des hypothèses sur des gènes qui peuvent servir des marqueurs de maladies [BSE+04]. Plus précisément, on a tendance à suivre le niveau d'expression des gènes en partant de trois objectifs : assigner de nouvelles fonctions biologiques aux gènes, inférer de nouveaux réseaux de régulation et/ou d'interactions, et identifier les gènes marqueurs pour le diagnostic ou le suivi de pathologies [Mic08].

A cet effet, des méthodes de classification de gènes ont été proposées pour identifier et regrouper les gènes qui partagent le même *profil d'expression*, c'est à dire un comportement transcriptomique similaire par rapport à un ensemble de conditions biologiques. Le profil d'expression d'un gène i est l'ensemble des valeurs des niveaux d'expression de ce gène par rapport à toutes les situations biologiques considérées : $Profil_i = \{M[i,k] \mid k = 1..p\}$.

Le critère de similarité (ou de distance) entre les niveaux d'expression (données numériques) est utilisé pour classifier les gènes. La distance euclidienne ou le coefficient de corrélation de *Pearson* sont souvent utilisés pour calculer la distance entre les niveaux d'expression. En partant du principe d'arbre phylogénique, un gène est regroupé avec le gène ayant une valeur d'expression la plus similaire, et au fur et à mesure d'autres gènes leur sont rajoutés en respectant le critère de similarité. Remarquons qu'au final, l'ensemble des gènes partageant le même profil d'expression

4. http ://sidr-dr.inist.fr/

est parfois appelé lui-même un profil d'expression.
Dans la Figure 4 (A) nous avons un exemple de mesure du niveau d'expression d'un gène (CCNE1) à des moments différents du cycle de division cellulaire. L'ensemble des valeurs d'expression constituent le profil d'expression de ce gène [DDG08]. Dans la partie (B) de la Figure nous avons un groupe de gènes qui suivent un profil d'expression similaire [LCH$^+$04].
Afin de pouvoir affecter des gènes aux profils d'expression, nous avons besoin d'un algorithme

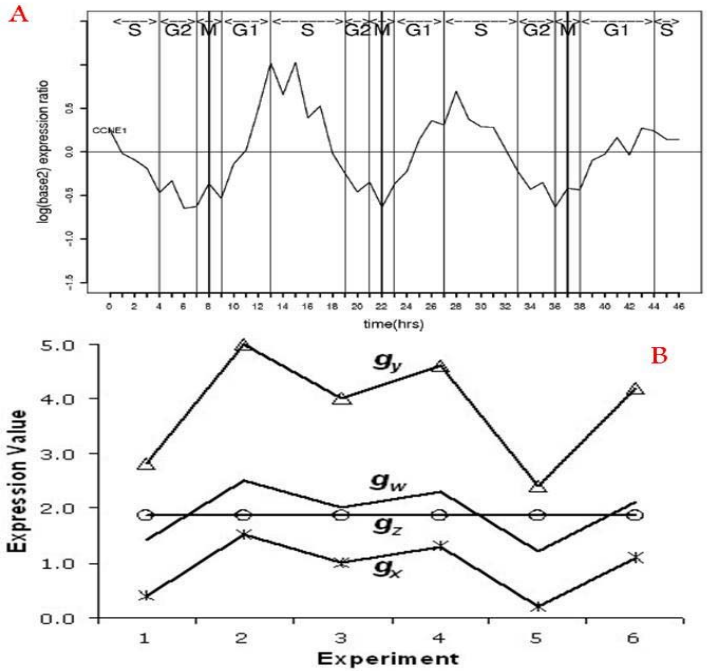


FIGURE 4 – Variation des niveaux d'expression de gènes par rapport à plusieurs situations biologiques. Partie (A) représente un profil d'expression du gène CCNE1, la partie (B) regroupe plusieurs gènes ayant des profils d'expression similaires [DDG08].

de classification [Bis06], et d'une métrique pour identifier les gènes qui ont un comportement transcriptomique similaire [LRB09a, CCT10, Cha07]. Soit Δ un ensemble de gènes, tel que : Δ=$\{G_i, i = 1..n\}$. Soit P une partition de Δ en m sous ensembles C_j tel que : $j = 1..m$. Ces ensembles sont appelés clusters (partitions) si et seulement ils satisfont ces conditions [TK06] :

$$\begin{cases} C_i \neq \emptyset & i = 1..m \\ \bigcup_{i=1}^{m} C_i = \Delta & \\ C_i \cap C_j = \emptyset & i \neq j (\textit{dans le cas de clusters non chevauchants}). \end{cases} \tag{3}$$

Plusieurs algorithmes de classification ont été utilisés afin de regrouper les gènes dans les profils d'expression, utilisant diverses métriques de similarité ou de distance (dissimilarité) pour comparer les niveaux d'expression [ACDI06, BDY99]. Les mesures de similarité sont utilisées pour identifier des paires de gènes dans Δ qui sont assez similaires en terme de niveaux d'expression. Soit Sim une fonction de similarité, $Sim(G_1, G_2)$ devient grande si les deux gènes G_1 et G_2 sont similaires, sinon cette valeur devient faible. Ayant deux gènes G_1 et G_2 cette mesure nécessite

les conditions suivantes :

$$\begin{cases} Sim(G_1, G_2) \in [0, 1] \\ Sim(G_1, G_1) = 1 \\ Sim(G_1, G_2) = Sim(G_2, G_1) \end{cases} \quad (4)$$

Dans le cas d'une distance la deuxième condition devient $Dist(G_1, G_1) = 0$.
Cependant, le choix de la métrique pendant le processus d'extraction de profils d'expression est crucial. Soit deux gènes G_1 et G_2, ayant respectivement les profils d'expression (sur p situations) : $Prof_{G_1}$=$(x_1, x_2, ..., x_p)$ et $Prof_{G_2}$=$(y_1, y_2, ..., y_p)$ respectivement, tel que : x_k=$M[1, k]$ et y_h=$M[2, h]$, avec M une matrice d'expression.
Parmi les métriques souvent utilisées sur les données d'expression, il y a la distance euclidienne δ_E qui se calcule par la différence des coordonnées des deux vecteurs représentant deux objets : $\delta_E(G_1, G_2)$=$\sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$. Une des limites qui concerne l'utilisation de la distance euclidienne sur les données d'expression s'illustre dans le cas de la taille de l'échelle de variation des niveaux d'expression, c'est à dire deux gènes qui sont présents dans un même profil peuvent avoir des amplitudes de variation de leurs niveaux d'expression dans les conditions biologiques très distantes [Qua01]. De plus cette mesure ignore les dépendances (ou relations) qui peuvent exister entre les niveaux d'expression [DDCG09]. Le coefficient de *Pearson* a aussi été utilisé par les méthodes de classification de gènes dans les profils, dont le principe est de vérifier si les valeurs d'expression de deux gènes évoluent dans le même sens (co-régulation). Ce coefficient de corrélation entre deux profils d'expression se définit par : τ_{G_1,G_2}=$\frac{\sum_{i=1}^{p}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{p}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{p}(y_i-\overline{y})^2}}$.
Ce coefficient suppose que les données sont linéaires, et est aussi sensible aux valeurs aberrantes [Meu05, Dra03]. Une autre métrique de distance couvrant à la fois la distance euclidienne et le coefficient de corrélation a été utilisée dans [DDCG09], pour fournir un indice de proximité en valeur en fonction de la proximité et de la forme de distribution des données.
Chacun des algorithmes de classification des données d'expression a des avantages et des inconvénients, selon qu'il permet de remédier aux problèmes de valeurs manquantes dans la matrice d'expression ou il respecte la multi-appartenance des gènes.
Dans la littérature on distingue deux catégories d'algorithmes de classification : Les méthodes supervisées (supervised learning) et non supervisées (unsupervised learning) [ACDI06, Slo02, LC03].

2.3.2 Extraction de profils d'expression par classification supervisée :

Ces approches se basent sur l'utilisation d'un expert pour classifier les données en plusieurs classes. En partant d'un jeu de données dit d'apprentissage, un modèle est créé (ensemble de règles) afin de prédire les classes adéquates des éléments d'un nouveau jeu de données. Les auteurs comme *Slonim et al*, présentent un schéma général d'un processus de classification supervisée de données d'expression [Slo02]. A partir de la matrice d'expression initiale, un sous-ensemble de gènes est choisi aléatoirement pour servir comme classe d'apprentissage. Pendant la phase de validation, on extrait de manière répétitive des échantillons de tests pour évaluer les performances du modèle créé.
L'algorithme des k plus proches voisins (KNN : K Nearest Neighbor), fait partie des techniques de classification supervisée les plus simples à implémenter. A partir d'un jeu de données d'apprentissage et d'une métrique (mesure de similarité ou distance), on décide l'appartenance d'un objet à une classe qui possèdent les $K - NN$ éléments qui lui sont les plus similaires par un vote majoritaire. Cet algorithme a été utilisé par [RAM$^+$08, LGM$^+$05, PJS$^+$10], pour prédire

des classes de profils d'expression relatives aux cancers en utilisant un ensemble de gènes déjà connus. L'algorithme nécessite deux paramètres : K le nombre représentant le nombre de voisins, et un deuxième paramètre appelé marge d'erreur, qui représente le seuil de tolérance acceptable. Un autre algorithme souvent utilisé dans le paradigme de classification supervisée sur les données d'expression, s'appelle la classification des centroïdes (NCC : Nearest Centroïd Classifier). Cette méthode a fait l'objet d'étude par *Tibshirani et al.* dans leurs travaux de référence sur l'outil PAM (Prediction Analysis of Microarrays) [THNC02]. Un ensemble de collections de classes de gène (profils d'expression) relatives aux cancers ont été préparées a priori, où chaque classe est étiquetée par un gène représentant le barycentre (moyenne des niveaux d'expression) de la classe. Pendant l'étape de prédiction, les gènes de l'échantillon en test sont classés par rapport à la distance de leur niveau d'expression avec les différents centroïdes des classes déjà préparées par l'expert. Un ensemble de 43 gènes, pouvant servir de marqueurs de diagnostic pour les tumeurs SRBCT (Small Blue Round Cell Tumors), ont pu être affectés à 4 profils d'expression déjà préparés et qui représentent les quatre catégories distinctes de diagnostic de ces tumeurs [THNC02]. D'autres techniques de classification supervisée ont été utilisées pour affecter des gènes dans les profils d'expression. Par exemple la classification linéaire discriminante (LDA : Linear Discriminant Analysis) qui suppose que les données sont paramétriques et suggère un hyperplan séparateur du jeux de données en deux classes. Le gène dont le profil est inconnu est projeté dans l'espace qui lui est proche [DFS02]. L'inconvénient de cette méthode est qu'elle est limitée à deux classes et donc deux profils d'expression avec les données des puces à ADN ce qui n'est pas réellement le cas en pratique. Cette limitation pourrait être dépassée en appliquant les machines à vecteur support (SVM : Support Vector Machine) [BGL$^+$00], qui suggèrent l'utilisation d'autant de vecteurs séparateurs que de dimensions du jeu de données d'apprentissage, notamment dans la version MC-SVM (Multi Category-SVM) comme présenté dans [SAT$^+$05]. Ces auteurs insistent sur le fait que les MC-SVM sont plus performants en terme de minimisation du taux d'erreur que les autres approches supervisées comme les K-NN, classification par centroïdes,... .
Le classification avec les réseaux de neurones [Ben06] a été utilisée par *Khan et al.* [KWR$^+$01] en entrainant un modèle neuronal pour classifier des gènes par rapport à des profils d'expression relatifs aux tumeurs SRBCT, comme dans [THNC02]. Pour des raison de complexité, ce modèle choisi dans cette étude est linéaire (pas de couche cachée de perceptrons). Néanmoins cette approche est peu utilisée du fait de la complexité et du phénomène de sur-apprentissage souvent observé à cause du rapport disproportionné des quantités des données d'expression disponibles pendant l'apprentissage et le test. Des approches de validation des résultats de l'apprentissage supervisé ont été introduites pour réduire l'erreur en test. En effet, le type de données manipulées souvent liées à l'étude des cancers nécessite une grande prudence et précision. Les approches comme la division du jeu de données en une partie pour l'apprentissage et une autre pour le test sont souvent utilisées. D'autres techniques de validation comme le *bootstraping*, *validation croisée*, *bagging* ont aussi fait l'objet d'études [ACDI06].
Bien que plusieurs travaux sur l'extraction de profils d'expression ont été réalisés avec les méthodes de classification supervisée, il faut admettre que la classification non supervisée est plus utilisée et dominante dans ce type de données [LH03]. De fait, la taille de ces données ne cesse de s'accroître et les connaissances du domaine sur l'expression des gènes sont de plus en plus complexes. Par conséquent il est très difficile de créer des hypothèse a priori sur le comportement des gènes du point de vue transcriptomique et donc de préparer des classes d'apprentissage.

2.3.3 Extraction de profils d'expression par classification non supervisée :

Les approches de ce paradigme de classification n'utilisent pas d'ensemble d'entrainement ou d'apprentissage, et n'ont aucune information a priori sur l'étiquetage des données à traiter. Le terme *clustering* a été introduit pour désigner le processus de regroupement automatique des données observées en plusieurs groupes distincts appelés *clusters* [Mac67].
Plusieurs méthodes de classification non supervisée ont été appliquées aux données d'expression [ACDI06]. Par exemple, *Eisen et al.* sont considérés parmi les premiers à avoir utilisé la technique de *clustering hiérarchique* pour "clusteriser" et visualiser les gènes de la levure selon leur profil d'expression [ESBB98]. Le principe général du clustering hiérarchique est l'utilisation d'une représentation phylogénique des objets. Il existe deux sous-catégories d'algorithme de clustering hiérarchique : clustering ascendant (*AGNES : Agglomerative Nesting Algorithm*) et descendant (*DIANA : Divisive Analysis Clustering*) [TK06]. Le *clustering hiérarchique ascendant* se base sur la fusion successive de clusters déjà existants. Le clustering hiérarchique ascendant considère au début tous les gènes comme étant un seul singleton, et fait appel à une matrice de similarité ou distance initiale (euclidienne, *Pearson* ,...). La première étape rassemble dans un cluster de départ les deux gènes les plus proches. Itérativement on répète ce processus en fusionnant à chaque étape les deux clusters les plus similaires. L'algorithme s'arrête quand il ne reste plus de gènes à fusionner. Au final le résultat peut être visualisé sous forme d'arbre binaire appelé *Dendrogram*. Avec n objets, on peut obtenir 2^{n-1} réarrangements linéaires consistants sous forme d'arbre binaire.
Le problème ici serait de savoir mesurer la distance entre groupes (clusters) de gènes (ou d'objets) pour pouvoir fusionner les gènes déjà affectés aux clusters (profils) avec le reste des gènes qui ne sont pas encore clusterisés. Pour cela plusieurs techniques, appelées règles d'agglomération, sont nécessaires pour calculer les différentes liaisons entre les clusters disjoints [Tol05].
Dans la Figure 6, on a un résumé des différentes mesures inter-clusters qui existent, utilisées par les algorithmes de clustering agglomératives. Le lien moyen (*average linkage*) appelé aussi

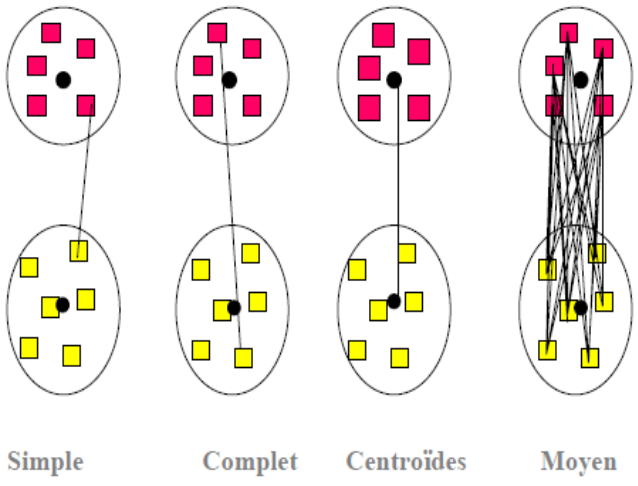


FIGURE 5 – Différentes mesures de distance inter-clusters utilisées par les algorithmes de clustering hiérarchique. Chaque cercle représente un cluster de gènes. Le lien ou distance entre clusters peut être : simple, complète, moyenne, centroïdes.

UPGMA (*Unweighted Pair Group Method of Aggregation*) est la distance inter-clusters la plus

utilisée de nos jours. Cette distance se calcule par la moyenne des distances entre toutes les paires de gènes de chaque cluster [SM58]. Le lien simple quant à lui (*SLINK : Single Linkage*) prend en considération la distance la plus petite entre deux gènes des deux clusters. C'est le même principe que celui du plus proche voisin [KR05]. Contrairement à cette distance, il existe le lien complet (*CLINK : Complete Linkage*) où l'on prend en considération les deux gènes les plus éloignés dans les deux clusters. D'autres distances comme les centroïdes (barycentre) appelé sous le nom de *Ward's Linkage* sont aussi utilisée en calculant la distance entre les centres de gravité des deux clusters.
Les étiquettes des gènes présents dans une matrice d'expression peuvent être réordonnées à la suite d'un clustering hiérarchique. Chaque cellule de la matrice d'expression peut être colorée avec une intensité de couleur représentant le niveau d'expression du gène qui s'exprime dans la condition adéquate. Cette matrice colorée s'appelle une *carte thermique* ou *Heatmap* [WF09], où le vert représente généralement un niveau d'expression bas, tandis que le rouge représente un gène fortement exprimé. *Eisen et al.* [ESBB98], comme présenté dans la Figure 6, ont utilisé le clustering hiérarchique avec une visualisation par Heatmap pour illustrer qu'un groupe de gènes se trouvant dans un même cluster et positionnés dans une même région de la Heatmap avec une intensité de couleur proche ou similaire, représente un groupe de gènes avec un profil d'expression similaire. A chaque groupe de gènes pourraient donc être associées des fonctions biologiques expliquant leur regroupement. Il existe d'autres variantes des Heatmap avec cette fois-ci deux dendrogrammes, un pour les lignes (gènes) et l'autre pour les colonnes (situations biologiques).
Le clustering hiérarchique descendant (top down) procède de façon inverse. L'approche considère l'ensemble des données de départ comme un cluster initial unique, et le divise en deux clusters descendants. Cette division doit vérifier la séparabilité entre les deux clusters créés, c'est à dire que ces derniers doivent être le plus distants que possible. Cette procédure est ensuite appliquée à chacun des descendants récursivement jusqu'à ce qu'il ne reste plus que des clusters ne contenant qu'un seul gène. Le résultat peut être visualisé grâce à des arbres binaires ascendants appelés Cladogrammes. Le clustering DIANA a été peu utilisé sur les données d'expression comparé à l'approche AGNES [AMP+10].
En plus du clustering hiérarchique, d'autres techniques de classification non supervisée de partitionnement ont été appliquées avec les données transcriptomiques dans le but d'affecter des gènes aux profils d'expression. Le principe général des techniques de partitionnement est de minimiser la distance intra-cluster tout en maximisant la distance inter-clusters pour un nombre k de clusters donné. La technique des $K-means$ est parmi les algorithmes de partitionnement les plus connues [Mac67]. Au départ, les gènes sont répartis autour de k centres choisis aléatoirement. A partir de ces k barycentres, chaque gène est affecté au centre dont le niveau d'expression est le plus proche formant un cluster. Le centre est mis à jour en calculant par exemple la moyenne des niveaux d'expression des gènes déjà présents dans chaque cluster. L'algorithme se réitère récursivement jusqu'à convergence c'est à dire lorsque l'on n'observe plus de grand changement des barycentres.
Une amélioration des $K-means$ a été proposée par *Kaufman et al.* dans [KR05] pour remédier aux problèmes de barycentres qui représentent une limitation des $K-means$. En effet, les auteurs dans la version PAM (*Partitionning Around Medoïds*) suggèrent de calculer la médiane de chaque cluster à la place du barycentre. Il est plus significatif d'utiliser le gène centroïde déjà présent dans le cluster comme centre que de calculer un barycentre et donc une position artificielle.
Les auteurs comme *Tavazoie et al.* [THC+99] ont utilisé le clustering $K-means$ pour regrouper des gènes de la levure synchronisés dans 15 points de mesure temporelles pour deux cycles cellulaires. Selon leurs résultats, ils ont identifié des groupes de gènes qui sont co-régulés. Cette

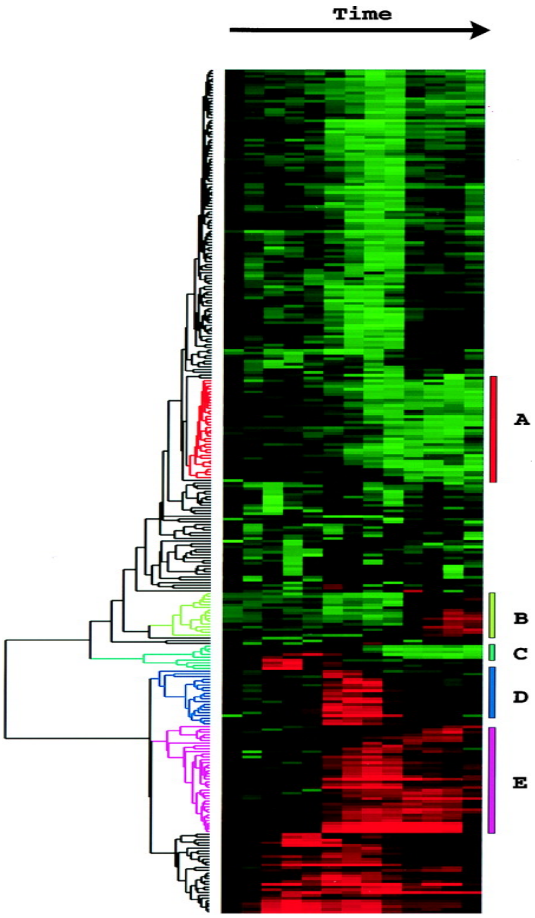


FIGURE 6 – Clustering hiérarchique et visualisation avec Heatmap. En ligne sont présents une liste de gènes de la levure, dont les étiquettes sont organisées avec un clustering hiérarchique et représentées avec un dendrogramme. Les colonnes représentent les situations biologiques (instants de mesure du niveau d'expression de chaque gène. Chaque cellule coloriée représente le niveau d'expression de la ligne (i) dans la situation (j). Après clustering hiérarchique des régions homogènes se forment donnant figure à un ensemble de gènes co-régulés et ayant un profil d'expression similaire. A chaque groupe de gènes pourraient être associées des fonctions biologiques expliquant leur regroupement [ESBB98].

étude a démontré que la méthode des $K-means$ a été capable de clusteriser des gènes dans des clusters biologiquement valides. Ainsi, cette approche pourrait être généralisée pour identifier des réseaux de régulation de gènes [DLS00].
Cependant, dans la nature réelle des données biologiques, un gène peut intervenir dans plusieurs processus biologiques à la fois. Par conséquence il serait plus logique de permettre à un gène d'appartenir à plus d'un profil d'expression à la fois. Pour cela des techniques de classification multi-appartenance ont été appliquées aux données d'expression. Parmi elles, l'algorithme $Fuzzy\ C-Means$ ou FCM [Bez81], prend en considération une fonction d'appartenance basée sur la logique floue, c'est à dire qu'à chaque gène du jeu de données est associé un poids représentant son degré d'appartenance aux clusters auxquels il appartient [WHJO03, DK03]. Après avoir utilisé le clustering hiérarchique, *Eisen et al.* dans un travail plus récent ont testé les FCM sur leur jeu de données transcriptomiques relatives à la levure [GE02]. Ce clustering chevauchant a permis de mettre en évidence des clusters de gènes co-régulés qui se recouvrent, et d'étudier la corrélation entre ces clusters et les sites de facteurs de transcription présents dans les promoteurs de gènes. Cette méthode est efficace notamment pour faire face au caractère bruité des données d'expression.
Les cartes topologiques de *Kohonen* ou cartes auto-organisatrices (*SOM : Self Organizing Maps*) [KHKL96, Leb03], sont considérées comme une extension des $k-means$. Elles ont été utilisées dans [TSM+99] pour clusteriser les données d'expression de 6000 gènes humain des organes hématopoïétiques responsables de la production de cellule sanguines, issus de quatre tissus différents (conditions biologiques). L'approche prend en considération un ensemble de nœuds organisés sous forme de graphe non dirigé (carte) et une distance entre ces nœuds. Au départ la position des nœuds est choisie aléatoirement. Leur position se réadapte après plusieurs itérations, pour couvrir l'ensemble du jeu de données, en rapprochant d'un nœud les gènes qui sont similaires dans la matrice d'expression. La complexité de la méthode réside dans l'étape de l'initialisation, du choix des paramètres, et de la structure topographique de la carte.
Le choix du nombre de clusters (k) à générer dans les algorithmes $K-means$ et FCM reste non trivial. En effet, l'avantage du clustering hiérarchique est qu'on a une vision globale de l'organisation des gènes en dendrogramme. Avec un bon niveau de découpage de cet arbre binaire, on obtient des clusters. Dans le cas des partitions, on est souvent face à des résultats non déterministes, où la position des centres change si on répète l'expérience plusieurs fois. Pour cela, on a souvent tendance à avoir réalisé au préalable un clustering hiérarchique pour avoir une distribution globale du jeu de données, et une indication sur le nombre k de clusters attendus. C'est ce nombre k qui peut alors être utilisé dans un algorithme de partitionnement chevauchant pour permettre à un gène d'appartenir à plus d'un cluster à la fois. Dans le cas où le clustering hiérarchique est appliqué sur les deux dimensions de la matrice d'expression, c'est à dire les gènes et les conditions biologiques, il faut faire attention à la nature de ces conditions biologiques. S'il s'agit d'une suite temporelle de mesures, on a souvent tendance à respecter l'ordre des points de mesure au lieu de les mélanger par le clustering hiérarchique [Mic08]. Enfin, d'autres analyses sont possibles en utilisant des méthodes factorielles dont le but est de réduire les dimensions de l'espace des données d'expression et donc la complexité du problème [RSA00].

2.3.4 Méthodes symboliques de fouille de données d'expression

Les données d'expression intéressent de nos jours de plus en plus de chercheurs à cause de leur mise à disposition publique et de la véracité des hypothèses qu'on peut en dériver. De ce fait, des techniques de fouille de données symboliques [Nap05] ont été appliquées aux matrices

d'expression.
Récemment, des algorithmes de bi-clustering, connus aussi sous le nom de bi-partition, co-clustering, subspace clustering, bi-dimensional clustering, two-way clustering, et block clustering, sont utilisés pour regrouper simultanément des objets partageant des attributs similaires [MO04, YG00]. Le bi-clustering regroupe les lignes et les colonnes de la matrice d'expression pour obtenir des sous matrices représentants des gènes co-régulés. Quelques implémentations de cet algorithme sont disponibles dans le package R bioconductor [5].
L'analyse formelle de concepts (AFC) [GW99], considérée comme un cas particulier du bi-clustering, a été appliquée avec succès sur des contextes booléens (ou binaires) en codant certaines propriétés d'expression des gènes afin de regrouper un ensemble de gènes et de situations biologiques [BBR$^+$07]. Soit K un contexte formel $K =(G, M, I)$ où G est un ensemble de gènes, M un ensemble de situations biologiques et I une relation binaire d'incidence entre G et M pour décrire qu'un gène d'une ligne i s'exprime dans la situation biologique j. Un concept formel issu de l'AFC est en fait un bi-ensemble de gènes et de situations tels que l'ensemble des gènes (extension du concept) présentent les mêmes propriétés d'expression dans l'ensemble de situations (intention du concept). En ce sens, l'AFC constitue un bon outil de classification conceptuelle. Ainsi *Blachon et al.* [BBR$^+$07] recherchent des groupes de gènes sur-exprimés dans un ensemble de situations en utilisant l'AFC. Bien que plus explicites dans leurs résultats, la principale difficulté des méthodes symboliques réside dans la discrétisation des mesures d'expression, nécessaire pour encoder de façon binaire les propriétés d'expression des gènes. Par exemple on dira que le gène est exprimé dans telle situation si son niveau d'expression est dans un certain intervalle de valeurs. Suivant la distribution des mesures et le mode de discrétisation on définira des intervalles d'expression faible, moyenne ou forte [KDN08]. Par ailleurs, le nombre important de groupements potentiels peut être contrôlé par le biais de contraintes telles que la cardinalité minimale de l'extension et de l'intention pour les concepts formels ou le support et la confiance minimum pour les règles d'association [BJ10]. Ces contraintes sont susceptibles d'améliorer la pertinence biologique des bi-ensembles obtenus.
Des auteurs comme *Pensa et al.* ont réalisé un processus d'extraction de motifs fréquents pour obtenir un ensemble de gènes co-exprimés dans certaines conditions [PB04]. En utilisant un contexte binaire comme le contexte K il est possible de récupérer les motifs fréquents, en isolant des attributs ayant un support prédéfini, c'est à dire un nombre d'objets minimum qui partagent ces attributs. Avec ces motifs, il est possible de produire des règles d'association. Ces méthodes sont disponibles dans la plateforme *Coron* [6] de l'équipe Orpailleur au Loria.
Les résultats du clustering (supervisé ou non) et des méthodes symboliques de fouille de données, nécessitent des techniques de validation, pour évaluer la qualité des regroupements obtenus. Plusieurs indices de qualité ont été proposés dans la littérature [GSU08, FL08, LAJS05, WY05, KLL04].

3 Prise en compte des connaissances du domaine dans l'analyse transcriptomique

3.1 Motivation et objectifs

Les méthodes d'extraction de profils d'expression prennent en entrée les données d'expression sans tenir compte des connaissances du domaine. Pourtant ces connaissances peuvent se révéler

5. http ://cran.r-project.org/web/packages/biclust/index.html
6. http ://coron.loria.fr/site/index.php

	Données numériques					Données symboliques			
	Condition 1	Condition 2	Condition 3	...	Condition p	Terme d'annotation 1	Terme d'annotation 2	...	Terme d'annotation K
Gene 1	314	445	90	1234	955	embryonic development	sensu Metazoa	...	gastrulation
Gene 2	78	123	515	63	1700	pseudocleavage	development	...	dorsal closure
Gene 3	15	46	789	103	500	cell killing	pigmentation	...	sulfur utilization
...	.	.	.	.	.	.	.	.	.
Gene i	.	.	.	.	.	.	.	.	.
....	...	...	...	...	...	...	...	...	...
Gene n	400	612	19	46	150	catalytic activity	morphogen activity	...	toxin transporter

TABLE 3 – Exemple de matrice pour la prise en compte des connaissances du domaines dans l'analyse de données transcriptomique. Les données numériques qui représentent les niveaux d'expression d'une liste de gènes sont complétées par des données symboliques qui représentent les connaissances du domaine.

fort utiles pour l'interprétation des données d'expression. Ces connaissances peuvent concerner d'une part les situations étudiées dans les expériences considérées, d'autre part les gènes impliqués dans le dispositif expérimental. Ainsi, par exemple, le fait de savoir que parmi les échantillons testés, certains proviennent de patients présentant telles ou telles variations génomiques peut aider à interpréter les profils d'expression observés. Du point de vue des gènes, la caractérisation fonctionnelle des profils d'expression consiste à analyser les annotations fonctionnelles des gènes présents dans chacun d'eux dans le but de comprendre pourquoi un ensemble de gènes se comporte de façon similaire pour un ensemble de situations donné [HDS+03a, ZFW+03, ENS+09]. Les connaissances du domaine sont donc ici représentées comme les annotations fonctionnelles des gènes ou des situations et peuvent être issues de plusieurs bases de données, par exemple : la base de voies métaboliques (pathways) KEGG [KEG], la base UNIPROT [uni] pour la description des séquences de protéines, la base PFAM [FMSB+] pour la description des familles de protéines, le vocabulaire d'annotation GO (Gene Ontology) [Con10], et de façon plus générale les ontologies bio-médicales. Le développement de ces ontologies est soutenu par deux organismes : NCBO (*National Center for Biomedical Ontologies*) et OBO (*Open Biological Ontologies*). Leur objectif est de regrouper dans un portail web des ontologies qui couvrent le domaine bio-médical. En effet, plusieurs ontologies sont en cours de développement et sont utilisées dans des domaines très variés [Bod09]. Par exemple, dans le domaine de la pharmacovigilance il y a les vocabulaires MedDRA (*Medical Dictionary for Drug Regulatory Activities*), WHO-ART (*World Health Organization Adverse Reaction Terminology*), qui sont utilisés pour coder les effets secondaires des médicaments [Mer08]. Le vocabulaire CheBI (*Chemical Entities of Biological Interest*) est utilisé comme ressource publique pour annoter les composantes chimiques des molécules [dMAD+10]. L'ontologie FMA (*The Foundational Model of Anatomy*) décrit de façon symbolique la structure du phénotype du corps humain [RM03]. Le vocabulaire SNOMED CT (*Systematized Nomenclature of Medicine-Clinical Terms*), est une collection de terminologies médicales couvrant différents secteurs des informations cliniques qui concernent les maladies [LHM93]. Le thésaurus COSTART (*Thesaurus of Adverse Reaction Terms*) est une ontologie regroupant aussi de façon structurée les informations concernant les effets secondaires des produits pharmaceutiques. Cependant, l'ontologie GO (*Gene Ontology*) est considérée comme la source d'annotations fonc-

tionnelles la plus utilisée pour annoter les produits de gènes et des protéines [Con10, AMI]. Plusieurs travaux en médecine ont fait référence à ces ontologies du fait de leur mise à disposition publique. Dans [WHM+09], les auteurs ont travaillé sur l'identification de gènes d'intérêt pour les maladies humaines en utilisant les annotations de plusieurs ontologies : GO, CHeBI, PATO .... De nombreux travaux existent comme [SSZ04, BW07, SFSZ05, AS04, BW07], où les auteurs ont utilisé les annotations GO pour réaliser une analyse fonctionnelle de gènes pour interpréter les résultats des puces à ADN. Cette analyse fonctionnelle suggère le regroupement des gènes sur la base de leur profils d'expression et de leur annotations fonctionnelles. Dans ce cas, les données d'expression sont considérées comme dans la Table 3. La même source d'annotation (GO) a été utilisée par *Schlicker et al.* dans [SLA10], pour classer les gènes selon leur similarité sémantique par rapport à une maladie d'intérêt accessible via la base de données OMIM [McK07]. Il est donc nécessaire ici de considérer plus particulièrement l'ontologie GO.

3.2 Annotation fonctionnelle des objets biologiques : l'ontologie GO

3.2.1 Notion d'annotation fonctionnelle et présentation de l'ontologie GO

Le vocabulaire structuré GO (Gene Ontology) [7] est considéré à juste titre comme l'un des plus importants projets en bioinformatique [ABB+00, Con10, BDH+09]. Cette ontologie est le fruit de la standardisation des nomenclatures des produits de gènes dans le but de rendre accessible la description des fonctions biologiques qui sont associées aux gènes des eucaryotes via un vocabulaire structuré. Dans la Figure 7 on constate l'évolution du nombre de citations du projet GO par rapport aux autres ontologies bio-médicales [JB10]. Dans la sous-section suivante nous allons présenter la structure de l'ontologie, le positionnement des annotations dans l'ontologie, comment est réalisé le processus d'annotation, et les aspects quantitatifs qui la concernent.

3.2.2 Aspects quantitatifs et structurels de GO

L'ontologie GO se compose d'environ 30 000 termes d'annotation organisés sous forme de vocabulaire contrôlé, décrivant les trois aspects des annotations fonctionnelles de gènes : Processus Biologiques (BP), Fonctions Moléculaires (MF) et Composants Cellulaires (CC) [ABB+00]. Cette ontologie est structurée sous forme de Graphe Acyclique Dirigé avec une racine unique (rDAG : rooted Directed Acyclic Graph), c'est-à-dire qu'il n'y a pas un chemin qui commence et se termine par le même nœud. Les nœuds du graphe de l'ontologie représentent les termes d'annotation c'est à dire les concepts, et les arcs représentent les relations sémantiques qui existent entre les annotations. Les trois aspects de l'ontologie (BP, MF, CC) constituent 3 sous-ontologies quasiment indépendantes de GO. Ils sont représentés sous forme de 3 sous-graphes (branches) orthogonaux. Il existe trois principales relations sémantiques dans cette ontologie : *is_ a*, *part_ of* et *regulates* [Con10]. Cette dernière se spécialise en deux sous-relations : *positively regulates* et *negatively regulates*. La subsomption "$terme_1$ is_a $terme_2$" signifie que le terme descendant $terme_1$ est une spécialisation du terme parent $terme_2$. La relation "$terme_1$ part_of $terme_2$" signifie que $terme_1$ est toujours partie de $terme_2$. Si $terme_2$ existe ceci ne signifie pas que $terme_1$ existe forcément. La dernière relation de régulation explique l'interaction de régulation entre deux processus biologiques. Elle peut aussi relier des termes de l'aspect MF à des termes de l'aspect BP. Cependant les deux premières relations *is_ a* et *part_ of* sont majoritaires dans cette ontologie. Nous avons récemment calculé sur la base de données de l'ontologie GO (version Octobre 2010) le pourcentage des différentes relations sémantiques, et 84% de ces relations sont de type *is_ a*.

7. http ://www.geneontology.org/

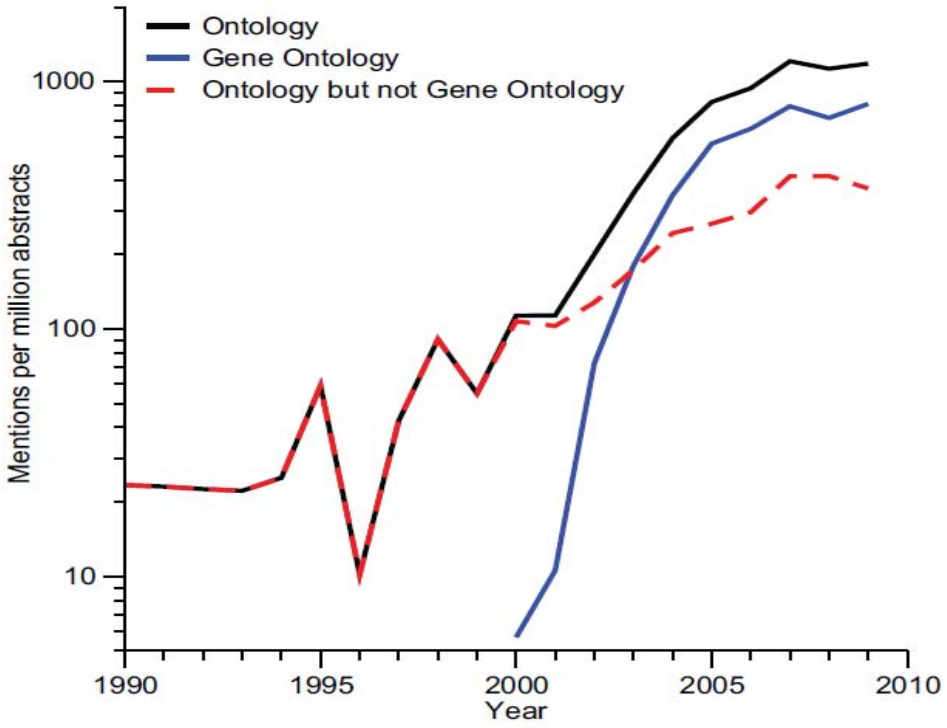


FIGURE 7 – Illustration du nombre de mentions de l'ontologie GO par million de résumés d'articles dans la base de données Medline. On constate son évolution exponentielle par rapport aux autres termes relatifs aux ontologies au sens général et plus spécifiquement aux ontologies bio-médicales [JB10].

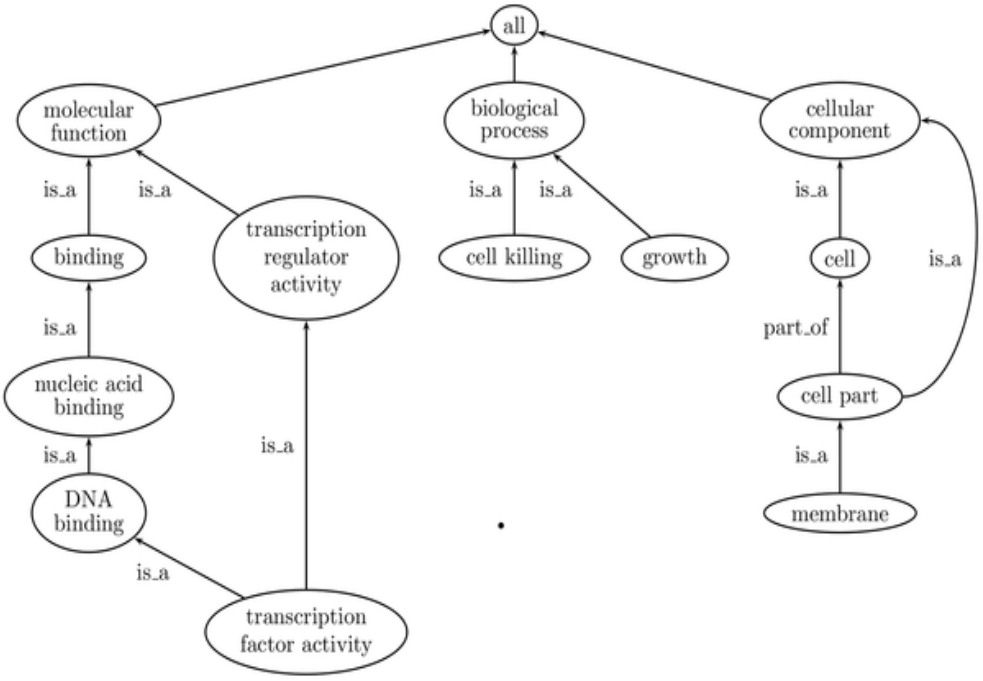


FIGURE 8 – Exemple de quelques termes d'annotation de l'ontologie GO. Chaque terme (concept de l'ontologie) est un nœud du rDAG. Les arcs représentent les relations sémantiques entre les annotations. Il existe trois sous-ontologies représentant les trois aspects de l'ontologie : Biological Process (BP), Molecular Function (MF), Cellular Component (CC). Ces trois aspects sont structurés via trois branches orthogonales du graphe réunis sous le même parent direct : le noeud Root=all [PFF+09].

Plus précisément, dans l'ontologie GO il y a 40700 entrées pour les relations *is_a* réparties en 3639 pour la branche CC, 10713 pour MF et 26348 pour la branche BP. La relation *part_of* est surtout présente dans la branche BP de l'ontologie avec 3239 entrées, puis la branche CC avec 933 entrées et encore loin devant la branche MF où il y a seulement 3 relations de type *part_of* reliant 3 paires de termes d'annotations : (*telomerase activity*, *template for synthesis of G-rich strand of telomer*), (*telomerase activity*, *telomeric template RNA reverse transcriptase activity*) et (*gap junction channel activity*, *gap junction hemi-channel activity*). Ce qui fait un total de 4175 relations de type *part_of* dans toute l'ontologie. La relation *regulates* est peu utilisée dans GO avec 1330 entrées seulement. Récemment une autre relation sémantique a été introduite, peu utilisée et qui s'appelle "*has_part*" en vue de représenter le rôle inverse à la relation *part_of* [Sch10]. Le consortium GO met à jour l'ontologie régulièrement en ajoutant de nouvelles annotations aux produits de gènes.

La Figure 8 illustre un exemple de quelques termes d'annotation dans le graphe de l'ontologie GO. On identifie trois aspects de l'ontologie GO. L'aspect MF (Molecular Function) rassemble les termes qui décrivent les activités biochimiques des produits des gènes (les protéines). Par exemple *transcription regulator activity* est une fonction moléculaire. L'aspect BP (Biological Process) regroupe les annotations qui sont des processus biologiques dans lesquels interviennent des gènes ou des produits de gènes. Le terme *Cell Killing* est un processus biologique. Dans le domaine de la biologie cellulaire on considère souvent un processus biologique comme étant une suite de fonctions moléculaires ordonnées. Le dernier aspect (CC), décrit les structures cellulaires

des organismes dans lesquelles on peut trouver le produit des gènes. Par exemple, une *membrane* est une partie de la cellule. Les trois aspect Biological Process, Molecular Function et Cellular Component ont comme parent direct le super nœud $Root = all$.
Comme dans n'importe quel vocabulaire structuré, la profondeur d'un terme d'annotation dans le graphe de l'ontologie est en lien avec la spécificité de ce terme. Plus un terme est profond c'est à dire plus sa distance au nœud racine est grande, plus il apporte de détails sémantiques et fonctionnels sur les produits de gènes. Par conséquence un terme descendant est plus spécifique que ses parents.
Le consortium GO a indexé les annotations par des identifiants appelés GO Identifier (GO_ID). Ainsi, chaque terme GO est identifié par un symbole (GO :xxxx). par exemple *transcription regulator activity*, est identifié dans l'ontologie GO par GO :0030528. Le portail web AMIGO [8] permet de récupérer en ligne pour chaque terme GO son identifiant. Parmi les autres informations qu'on peut récupérer, sont le nombre et la liste des gènes et produits de gènes qu'annote ce terme d'annotation GO.
Chaque nœud dans le graphe de l'ontologie peut être connecté à plusieurs nœuds parents et nœuds descendants. Un terme d'annotation peut annoter un ou plusieurs produits de gène. Si un gène est annoté par un terme t_i par transitivité il l'est aussi par tous les parents de ce terme jusqu'à la racine selon la relation *is_ a*. C'est une contrainte imposée par le consortium GO, et qui s'appelle la "*True Path Rule*" [ABB$^+$00].
Les termes GO sont utilisés pour annoter des produits de gènes.
Le processus d'annotation, c'est à dire l'association d'un terme à un produit de gène (*mapping*), est supervisé par des indices de qualité. Ces indices s'appellent les codes d'évidence (EC) [RBH09]. Au total il y a 17 codes d'évidence, répartis en cinq catégories (Table 4). Seul le code d'évidence *Inferred from Electronic Annotation* (IEA) n'est pas assigné par un curateur parmi tous les autres codes d'évidence.
Les EC assignés de manière manuelle se répartissent en quatre catégories générales : Author statement, Experimental analysis, Computational analysis, et Curatorial statements. La catégorie "Author statement" signifie que les annotations ont été publiées par des auteurs dans une référence comme source d'information (TAS : Traceable Author Statement) ou non (NAS : Non traceable Author Statement). Une annotation du type EXP (Inferred from Experiment) signifie qu'elle provient d'une expérience dans un laboratoire. La catégorie "Experimental" regroupe six codes d'évidence (IDA, IPI, IMP, IGI, EXP et IEP). L'utilisation des codes d'évidence de cette catégorie signifie que l'annotation est faite par référence à une publication. La troisième catégorie "Computational Analysis" signifie que les annotations sont basées sur une analyse automatique supervisée par un expert curateur. Elle contient six types de codes d'évidence (ISS, RCA, ISA, ISO, ISM et IGC). Le code d'évidence IC (Inferred by Curator) fait partie de la catégorie "curatorial statement". Il est utilisé quand une annotation est inférée par un curateur à partir d'autres annotations GO qui ont déjà des codes d'évidence. L'autre code d'évidence appelé ND (No biological Data available) fait aussi partie de cette catégorie. Il signifie que le curateur n'a trouvé aucune information à propos de l'annotation. En pratique, le processus d'annotation suit un arbre de décision pour qualifier et superviser l'association d'un terme GO à un gène ou un produit de gène [evi]. Un gène peut donc être annoté par plusieurs termes GO, ou par un même terme GO mais avec des codes d'évidence différents.
D'après l'article de revue publié par *Rhee et al.*, il existait en 2008 plus de 16 millions d'annotations reliant des produits de gène et des termes d'annotation GO [RWDD08]. De plus environ 95% de ces annotations sont inférées de manière électronique, c'est à dire associées au code d'évi-

8. http ://amigo.geneontology.org/

Catégorie	Code d'évidence (EC)	signification	Nombre d'annotations associées à cet EC
Author statement	TAS	Traceable Author Statement	44 564
	NAS	Non-traceable Author Statement	25 565
Experimental	EXP	Inferred from Experiment	ND
	IDA	Inferred from Direct Assay	71 050
	IPI	Inferred from Physical Interaction	17 043
	IMP	Inferred from Mutant Phenotype	61 549
	IGI	Inferred from Genetic Interaction	8 311
Computational Analysis	IEP	Inferred from Expression Pattern	4 590
	ISS	Inferred from Sequence Similarity	196 643
	RCA	Inferred from Reviewed Computational Analysis	103792
	ISA	Inferred from Sequence Alignment	ND
	ISO	Inferred from Sequence Orthology	ND
	ISM	Inferred from Sequence Model	ND
	IGC	Inferred from Genomic Context	4
Curator statement	IC	Inferred from Curator	5 167
	ND	No biological Data available	132 192
Automatically assigned	IEA	Inferred from Electronic Annotation	15 687 382

TABLE 4 – Liste des codes d'évidence utilisés dans GO. Chaque annotation qui associe un produit de gène annoté à un terme d'annotation de GO est tracée par un code d'évidence de cette liste. La troisième colonne de ce tableau représente le nombre d'annotations qui sont associées à chaque code d'évidence. Ces données ont été publiées en 2008 [RWDD08].

dence IEA (Table 4). Cette observation a été par la suite confirmée en 2010 par *Schlicker et al.* dans [Sch10].

Plusieurs organismes de recherche se sont intéressés aux annotations GO. Ainsi des projets ont été fondés pour enrichir le processus d'annotation de nouvelles espèces. Par exemple, le projet (GOA) (*Gene Ontology Annotation Project*) a été fondé par l'EBI (European Bioinformatics Institute) et le groupe SWISSPROT. Grâce à ce projet plusieurs espèces ont été annotées soit partiellement, soit complètement (Table 5).
Dans ce tableau on peut voir qu'en 2008 la majorité des espèces n'ont pas un génome totalement annoté. Excepté les deux espèces *Schizosaccharomyces pombe* (levure à fission, *NCBI TAXON ID=4896*) avec 4930 gènes annotés, et *Saccharomyces cerevisiae* (Levure de bière *NCBI TAXON ID=4932*) avec 5794 gènes annotés, la moyenne du pourcentage d'annotation pour les autres espèces reste aux alentours de 63%. L'espèce Human (*NCBI TAXON ID=9606*) a un génome de 20877 gènes, qui sont annotés à 81.5%.

Les annotations GO sont utilisées dans de nombreuses applications pour les technologies des puces à ADN. Elles sont devenues un standard pour annoter les génomes de différentes espèces [Meu05]. Cependant, l'ontologie GO comporte quelques limites notamment en ce qui concerne la mise à jour des annotations. En effet, ce processus s'exécute souvent en ajoutant de nouvelles annotations, dans le but de corriger les erreurs et de maintenir la consistance de l'ontologie [Sch10]. Si des annotations sont supprimées de l'ontologie, elles seront marquées avec l'étiquette 'obsolète'. Ainsi le développement d'applications qui exploitent ce genre d'annotations doit être actualisé régulièrement, ce qui représente une tâche difficile. Pour faire face à ce problème, un nouveau projet a été créé et qui s'appelle Gene Ontology Next Generation project (GONG), qui suggère l'utilisation de langages de représentation de connaissance pour assurer une meilleur cohérence et faciliter l'évolution des applications [WSGA03]. Parmi ces langages il y a DAML+OIL (DARPA Agent Markup Language + Ontology Inference Layer) devenu il y a quelques années

Espèce (NCBI TAXON ID)	Nombre total de gènes	Pourcentage d'annotation
Schizosaccharomyces pombe (4896)	4 930	100%
Saccharomyces cerevisiae (4932)	5 794	100%
Mouse (10090)	27 289	67.4%
Caenorhabditis elegans (6239)	20 163	70.2%
Human (9606)	20 887	81.5%
Arabidopsis thaliana (3702)	27 029	98.5%
Rat (10116)	17 993	95.8%
Fruitfly (7227)	14 141	67.6%
Candida albicans (5476)	6 166	60.9%
Pseudomonas aeruginosa PAO1 (208964)	5 568	45.0%
Slime mold (44689)	13 625	50.6%
Trypanosoma brucei (5691)	9 154	42.8%
Zebrafish (7955)	21 322	63.7%
Plasmodium falciparum (5833)	5420	59.8%
Rice (39947)	41 908	71.3%
Chicken (9031)	16 737	36.2%
Cow (9913)	21 756	39.2%

TABLE 5 – Distribution des annotations de l'ontologie GO sur différentes espèces. Pour chaque espèce identifiée par un identifiant unique fourni par le NCBI : *National Center for Biotechnology Information* (http ://www.ncbi.nlm.nih.gov/) appelé NCBI TAXON ID, il y a le nombre de gènes présents dans cette espèce et le pourcentage de l'annotation (2008) [RWDD08].

OWL (Ontologies Web Language) [9], qui permet une classification de concepts tout en vérifiant la cohérence de l'ontologie.
L'ontologie GO est aussi présente sous d'autres formats, avec des outils qui permettent une bonne visualisation du graphe de l'ontologie. Parmi ces formats, il y a le standard OBO (Open Biological Ontologies) qui a comme objectif le regroupement de plusieurs ontologies en une seule plateforme web. Citons également de manière plus classique le langage XML, et les tables relationnelles.

4 Analyse fonctionnelle d'ensembles de gènes

4.1 Notion d'analyse fonctionnelle

L'analyse fonctionnelle de gènes s'appuie largement sur ce qui est maintenant utilisé dans divers domaines d'application de la biologie moléculaire, comme la prédiction de fonctions de gènes [KFD+03], la construction de réseaux métaboliques [DBD+04, ENB+07], la classification fonctionnelle de gènes [ZFW+03, ASAD+07, SSZ04], le processus de mapping d'ontologies biomédicales [RGF], et les mesures de similarité sémantique [PFF+09].
L'interprétation des données de puces à ADN représente l'exemple typique de l'utilisation des annotations GO comme source de connaissances pour l'analyse de données [KD05]. Le processus d'analyse fonctionnelle (Gene Functional Profiling) sert à la mise en évidence de clusters fonctionnels (de fonctions biologiques) représentés dans les profils d'expression. En effet, après l'extraction de profils de gènes différentiellement exprimés, une étude fonctionnelle de ces groupes de gènes est menée pour identifier les gènes qui partagent des fonctions biologiques similaires représentées par des annotations GO. Cette analyse fonctionnelle se base sur l'hypothèse communément admise chez les biologistes et qui suppose que les gènes co-exprimés partagent souvent des fonctions similaires [HP06].

9. http ://www.w3.org/TR/owl-features/

Pour réaliser une analyse fonctionnelle sur les données d'expression, on distingue deux approches différentes. La première approche prend en considération les termes d'annotations GO qui annotent l'ensemble des gènes étudiés comme étant un *sac de termes*. Ensuite, une *analyse d'enrichissement* des termes d'annotations est réalisée en utilisant des tests statistiques, afin d'identifier dans ce sac à termes, les termes GO qui sont les plus enrichis c'est à dire ceux dont la présence dans le sac à termes est le moins prévisible compte-tenu de leur fréquence d'utilisation en tant qu'annotation [ENS+09, ASAD+07, BS04b]. Ces fonctions enrichies décrivent alors les propriétés fonctionnelles de l'ensemble des gènes étudiés. Bien que l'étude d'enrichissement des termes GO soit la plus dominante dans les outils qui ont été proposés jusqu'à ce jour, une autre approche d'analyse fonctionnelle de données d'expression a été proposée, qui suggère l'utilisation d'une mesure de similarité sémantique pour réaliser un *classification fonctionnelle* de gènes.

4.2 Notion d'enrichissement en termes GO

Les méthodes d'analyse d'enrichissement offrent une interprétation statistique des résultats des données d'expression en identifiant les fonctions biologiques qui sont sur-représentées dans un groupe de gènes par exemple un profil d'expression ou autre [XGZD09].
Cette analyse est exécutée en 2 étapes : premièrement une liste de gènes d'intérêt est sélectionnée, par exemple un profil d'expression dans le cas des expériences de puces à ADN. On construit alors la liste des annotations GO de ces gènes. Puis un test statistique est appliqué sur les termes GO présents dans cette liste par rapport aux termes GO des gènes de référence, qui peuvent par exemple être tous les gènes d'une espèce donnée [KD05].

La probabilité d'occurrence d'un terme dans une liste d'annotations de gènes est souvent modélisée par la distribution hypergéométrique [SHH+06]. Cependant, si la liste de gènes est grande, cette distribution tend vers une loi binomiale. Avec la distribution hypergéométrique, on a souvent tendance à répondre à cette question : " *Dans une liste donnée de gènes d'intérêt (par exemple pour un profil d'expression), y-a-t-il un terme GO qui ne serait pas là, si le tirage des termes avait été fait par hasard seulement ?* ".
Cette question est la question de base du problème d'enrichissement en termes GO.
Plusieurs outils ont été proposés pour ce type d'analyse. Une liste non-exhaustive est répertoriée par le consortium GO sur ce lien : http ://www.geneontology.org/GO.tools.microarray. Les grandes différences résident dans le type du test statistique utilisé : *hypergéométrique*, *Khi2*, *binomial*, test exact de *Fisher*. Généralement les résultats obtenus avec ces outils se présentent, pour un ensemble de gènes d'intérêt, comme une liste de termes GO, classés par ordre croissant de leur valeur de significativité ou *p_value(terme GO)*. Plus la *p_value* d'un terme est faible, plus la liste est dite "enrichie" en ce terme, et plus la présence de ce terme dans la liste est significative. La *p_value* d'un terme GO dans une liste représente en effet la probabilité d'obtenir ce terme dans cette liste si l'on effectuait un tirage au hasard parmi tous les termes GO de l'ensemble de référence. Souvent un seuil est choisi pour ne garder que les termes GO ayant des *p_value* inférieures à cette valeur de seuil [BWG+04].
La Figure 9 présente quelques outils disponibles pour l'enrichissement des annotations. Par exemple, Onto-Express fut parmi les premiers à être proposé par *Draghichi et al.* dans le but d'interpréter les résultats issus des puces à ADN en extrayant les profils d'expression de gènes et les études d'enrichissement fonctionnel des annotations de ces gènes [KDOK02]. Un autre outil très similaire à Onto-Express est l'outil FatiGO [ASDUD04], disponible sur un portail web, mais qui a l'inconvénient de ne pas s'appuyer sur des tests statistiques puissants mais seulement sur l'étude du pourcentage de la distribution des annotations par rapport aux gènes. Sur le même

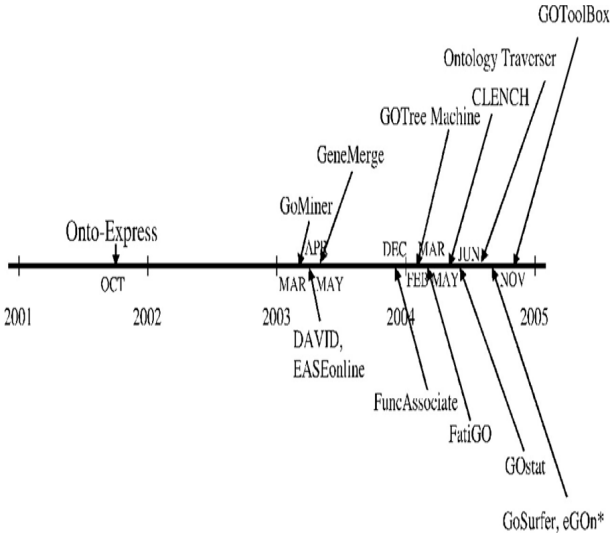


FIGURE 9 – Classement des outils d'enrichissement des annotations GO, par année d'apparition. Ces outils utilisent différentes approches statistiques pour générer les résultats d'enrichissement [KD05].

	liste globale de gènes	liste de gènes d'intérêt	
Annotés par (x)	X	Z	$\sum_{1,*}$
Non annotés par (x)	$N - X$	$T - Z$	$\sum_{2,*}$
	$\sum_{*,1}$	$\sum_{*,2}$	Total : $\sum_{*,*}$

TABLE 6 – Table de contingence deux à deux pour des gènes annotés ou non par un terme d'annotation GO (x). Cette table d'hypothèse est utilisée pour calculer l'enrichissement de ce terme dans la liste biologique considérée.

principe, des outils comme DynGO [LHW05], GOFish [BWKR03], Top-GO [YCCK] offrent une bonne visualisation du graphe des annotations mais sans indiquer leur niveau de significativité. Un peu plus tard, GOMiner a été proposé sous forme d'interface graphique Java (Ge'r(UI) [ZFW$^+$03]. L'outil est considéré parmi les premiers à avoir intégré un modèle statistique, en utilisant le test de *Fisher* à deux hypothèses. Soit N une liste globale de gènes différentiellement exprimés. On pose X le nombre de gènes à partir de N ayant un terme GO ($function_x$) comme annotation. Soit T le nombre de gènes dans une liste d'intérêt extraite à partir de la liste globale selon un critère prédéfini, par exemple selon leur appartenance à un profil d'expression. Soit Z le nombre de gènes de la liste d'intérêt qui ont le terme GO ($function_x$) comme annotation. On considère la table de contingence suivante (Table (6)) : Pour un terme GO ($function_x$) on a la table de contingence Table (6) avec les quatre paramètres ($\sum_{1,*}, \sum_{2,*}, \sum_{*,1}, \sum_{*,2}$) qui seront utilisés pour calculer la valeur de significativité de ce terme GO, via le test exact de *Fisher* sous l'hypothèse nulle :

$$p\_value(function_x) = \frac{\sum_{1,*}!.\sum_{2,*}!.\sum_{*,1}!.\sum_{*,2}!}{\sum_{*,*}!.X!.(N-X)!.Z!.(T-Z)!}. \quad (5)$$

D'autres outils utilisent le teste de *Khi2* en calculant l'enrichissement d'un terme GO par :

$$p\_value(function_x) = \frac{\sum_{*,*} \cdot (|\sum_{1,1} \cdot \sum_{2,2} - \sum_{1,2} \cdot \sum_{2,1}| - \frac{\sum_{*,*}}{2})^2}{\sum_{1,*} \cdot \sum_{2,*} \cdot \sum_{*,1} \cdot \sum_{*,2}}. \quad (6)$$

Ce test est souvent conseillé dans le cas d'une grande liste de gènes en entrée. Un exemple typique d'outils utilisant ce test est l'outil GO :Stat [BS04a].
Cependant la distribution hypergéométrique reste la plus utilisée pour estimer l'importance des annotations. La probabilité d'un terme GO selon cette distribution se calcule par :

$$p\_value(function_x) = \frac{\binom{T}{Z} \cdot \binom{N-T}{X-Z}}{\binom{N}{X}}. \quad (7)$$

Parmi les outils qui utilisent la distribution hypergéométrique pour calculer l'enrichissement des annotations il y a : GO : :TermFinder [BWG+04], GOrilla [ENS+09], GOlem [SHH+06], Ease Online [HDS+03b], GOToolBox [MBR+04] qui intègre aussi une fonction de clustering des annotations en utilisant la mesure de Czekanowski-Dice, GO Explorer (GOEx) [CFC+09], GObar [LKS05], ReviGO [SBSS11].
Les deux outils récents, à savoir GOrilla et ReviGO, offrent à notre connaissance la visualisation la plus complète des résultats d'enrichissement d'une liste de gènes. En effet, pour l'outil GOrilla, ayant une liste de gènes différentiellement exprimés en entrée, la liste de termes GO qui les annotent est extraite, et le processus d'enrichissement est lancé. Dès lors, les outils génèrent un sous-DAG de l'ontologie en ne gardant seulement que les termes GO considérés qui annotent les gènes étudiés. De plus, une intensité de couleur est associée à chaque nœud du DAG pour illustrer le niveau de sa *p_value*. Plus cette valeur est faible plus le terme GO correspondant est important, et donc plus la fonction biologique qu'il représente pourrait intervenir dans le processus de co-régulation des gènes. Dans la Figure 10, on a un exemple de visualisation des résultats obtenus avec ReviGO.
Le TreeMap généré avec l'outil ReviGO sert à identifier les catégories de termes GO qui existent dans l'ensemble d'annotations des gènes étudiés en entrée. Cette technologie de visualisation utilise un algorithme de clustering fonctionnel des termes GO en se basant sur leur similarité sémantique dans le graphe de l'ontologie. Un exemple de mesure de similarité sémantique utilisée est celle proposée par *Resnick et al.* (voir sous-section suivante pour plus de détails) [Res95]. L'intérêt d'utiliser des outils comme ReviGO dans ce genre d'analyse à partir de données d'expression réside dans la puissance de visualisation et l'interactivité des graphiques proposés.
Un autre outil qui s'appelle DAVID (Database for Annotation, Visualization and Integrated Discovery) [10] a été proposé. Il intègre plusieurs sources de connaissances appartenant à de multiples bases de données publiques : des annotations GO, des Pathways, des domaines de Protèines [SHT+07, HSL09]. L'outil DAVID constitue maintenant l'état de l'art en ce qui concerne l'analyse fonctionnelle des données biologiques. Il intègre plusieurs fonctionnalités, par exemple le *Gene Conversion Tool* qui permet la conversion des symboles de gènes vers plusieurs identifients (NCBI, Affymetrix ID,...), la fonction *Functional Annotation Tool* qui étudie l'enrichissement des annotations en utilisant la distribution hypergéométrique.
La fonctionnalité *DAVID Functional Classification tool* [HST+07] est de loin la plus réputée. Elle intègre un algorithme de clustering fonctionnel, pour regrouper des gènes ou des termes

10. http ://david.abcc.ncifcrf.gov/

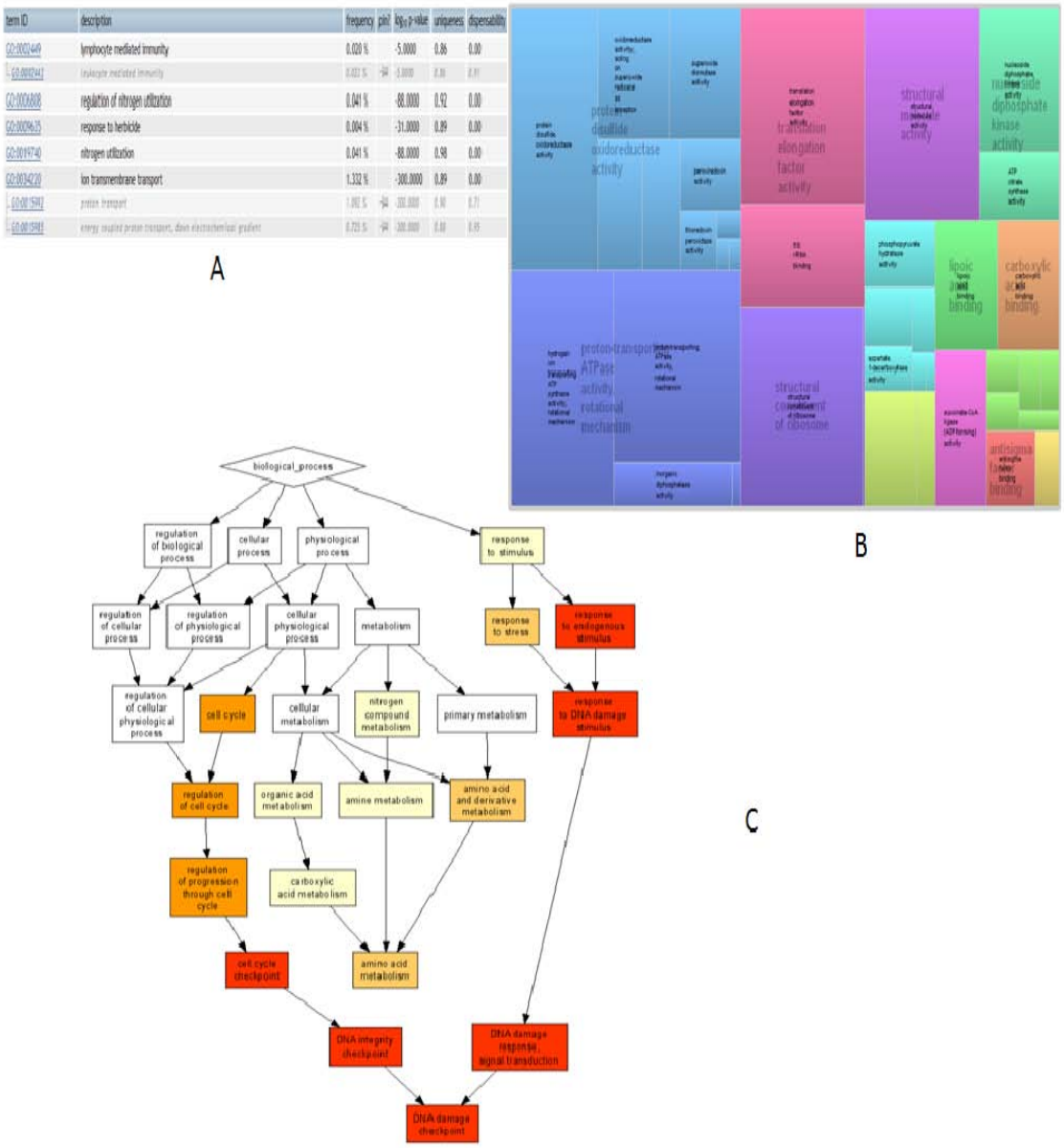


FIGURE 10 – Exemple de visualisation obtenue avec l'outil ReviGO, destiné à l'étude d'enrichissement des annotations. Un ensemble de termes GO qui annotent les gènes en entrée est trié selon les valeurs des p_value (Partie A de la Figure). Un TreeMap représentant le résultat du clustering fonctionnel des termes d'annotation est généré pour voir les catégories des termes d'annotations (Partie B de la Figure). Un sous-graphe de l'ontologie GO est généré en gardant les nœuds qui correspondent aux termes GO qui annotent l'ensemble de gènes étudiés (Partie C de la Figure). Une intensité de couleur est affectée à chaque nœud pour illustrer la valeur de sa p_value [SBSS11].

GO les plus similaires dans des groupes distincts. L'outil *DAVID Functional Classification tool* utilise la similarité basée sur la statistique Kappa [WC05]. Le principe de la similarité Kappa et du clustering fonctionnel de l'outil DAVID sera abordé dans une prochaine sous-section, car il servira d'outil de référence pour évaluer notre travail.
Une des limitations des approches d'enrichissement individuel des termes GO concerne leur dépendance vis à vis du choix de l'ensemble de gènes d'intérêts considéré. En effet, si cet ensemble de gènes change, en ajoutant ou en supprimant des gènes, le sac de termes qui annotent le nouvel ensemble de gènes va automatiquement changer. Ce qui a pour conséquence l'identification de nouveaux termes GO qui sont enrichis dans ce nouvel ensemble de gènes. Pour cela d'autres approches d'enrichissement plus globales ont été proposées (GSEA : Gene Set Enrichment Analysis) en prenant tous les gènes sans sous-échantillonnage [STM$^+$05]. De plus le choix du seuil de considération des *p_ value* reste un problème discutable. Il est difficile de choisir une valeur limite au dessous de la quelle on considère que l'enrichissement des annotations est significatif. Il n'existe aucune heuristique concernant ce seuil de *p_ value* qui est déterminé généralement par un jugement de l'expert.
Un autre problème réside dans le choix de la liste de gènes de référence (background) pendant le processus de calcul de l'enrichissement. En effet, certains outils d'analyse (GOToolBox, GOstat, GOMiner, FatiGO, GOTM), utilisent la totalité des gènes d'un génome comme ensemble de référence. Ce cas de figure peut devenir pénalisant dans le cas où on utilise les gènes d'une puce à ADN et que certains gènes du génome ne sont pas présents dans la puce. D'autres fluctuations issues des tests statistiques ont été observées, et des corrections ont été utilisées pour remédier à ces problèmes [BY01].
Puisque l'ontologie GO est organisée sous forme de rDAG, un problème de biais a été remarqué à cause des redondances des termes dues aux relations d'héritage multiple dans l'ensemble d'annotations traitées. Des solutions ont été proposées par *Alexia et al.* dans leur outil TopGO tool en incluant des connaissances a priori sur la structure du graphe de GO [ARL06]. Dans cette approche, un sac de termes est créé avec les annotations des gènes de la liste d'intérêt, puis un parcours dans le graphe de GO est effectué de bas en haut, pour chaque nœud (terme) visité, on calcule sa *p_ value*, si cette valeur est inférieure à un certain seuil prédéfini, alors on garde ce terme et on passe aux termes suivants, sinon on élimine ce terme du sac à termes, et par conséquent les gènes annotés par ce terme vont automatiquement le perdre. Il parait évident que le nombre de gènes dans la liste d'intérêt pourrait changer en éliminant les gènes qui n'avaient que les annotations avec forte *p_ value* (donc éliminées pendant le parcours). Ce processus pourrait être considéré comme une étape de nettoyage de la liste des gènes d'intérêt et du sac à termes qui les annotent. Les termes avec une forte *p_ value* (moins spécifique) sont exclus du processus d'enrichissement, ce qui limite le biais généré à cause de leur présence dans l'ensemble du sac à termes. *Grossman et al.*, ont proposé un outil similaire (Ontologizer tool) mais en tenant compte des relations parents/descendants [GBRV07]. Ces méthodes sont souvent appelées "parent-child enrichment approaches".
Mis à part ces deux derniers outils, la plupart des approches décrites pour mesurer l'enrichissement des termes d'annotations ne prennent pas en considération les liens sémantiques qui existent entre les annotations. Pourtant, et comme nous venons de le voir pour le cas de l'ontologie GO, il y a une sémantique très riche associée aux termes GO, qui s'illustre avec les relations sémantique notamment la relation *is_ a*. Il est regrettable que les outils proposés n'aient pas pris en considération ce point essentiel.

4.3 Notion de similarité sémantique sur GO

4.3.1 Définition générale des mesures de similarité

Dans les résultats d'analyse de données d'expression, les biologistes ont souvent tendance à vérifier si les gènes co-exprimés, partagent aussi les mêmes processus biologiques. En partant de cette hypothèse, une alternative aux études d'enrichissement des termes d'annotations consiste à calculer la similarité entre des produits de gènes annotés avec des termes GO. Pour cela, l'intérêt d'avoir une métrique pour mesurer la similarité entre des objets complexes (produits de gène), qui sont annotés par des annotations sémantiques (termes GO par exemple), se révèle impératif. La notion générale d'une similarité est de pouvoir mesurer des objets s'ils partagent des attributs ou caractéristiques similaires. En effet, dans le monde réel chaque objet peut être décrit par une liste d'attributs et de primitives le distinguant des autres objets. Pour cela, *Tversky et al.* définissent une mesure de similarité comme étant une fonction qui permet de comparer les attributs qui représentent les deux objets comparés [Tve77]. Les applications des mesures de similarité sont diverses, notamment en clustering où on a comme objectif de regrouper les objets qui sont similaires. Selon *Rifqi et al.*, les mesures de similarité peuvent être appliquées à différents types de données : par exemple binaire, numérique et structurée [LRB09b].
Les attributs sont binaires dans le sens où ils sont notés par présence (noté par 1) ou absence (noté par 0) dans les données. Dans ce cas, la description des objets est dans l'ensemble χ=$\{0,1\}^p$, où p est le nombre d'attributs. Un exemple réel dans le cas de données transcriptomiques est le marquage de présence ou non d'un terme d'annotation GO dans un vecteur définissant un gène. De façon générale, si l'on considère deux objets $x = (x_1, x_2, .., x_p)$ et $y = (y_1, y_2, .., y_p)$ dont les attributs sont dans χ^2, la similarité entre les objets x et y ayant des attributs binaires, se définit ainsi par une fonction :

$$SIM(x,y) : \chi^2 \longrightarrow \mathbf{R}, \quad verifiant\ 3\ conditions : \begin{cases} Positive : \forall x, y \in \chi, & SIM(x,y) \geq 0 \\ Maximale : \forall x, y \in \chi, & SIM(x,x) \geq SIM(x,y) \\ Symetrique : \forall x, y \in \chi, & SIM(x,y) = SIM(y,x) \end{cases}. \tag{8}$$

On définit pour chaque objet x et y les ensembles de leurs attributs présents dénotés $X = \{i \mid x_i = 1\}$ et $Y = \{i \mid y_i = 1\}$ respectivement. Le nombre d'attributs partagés entre les deux objets x et y est dénoté par : $a = |X \cap Y|$. Le nombre d'attributs dans x et non dans y est dénoté par : $b = |X - Y|$. Réciproquement $c = |Y - X|$ est le nombre d'attributs dans l'objet y et pas dans x. La quantité $d = |\overline{X} \cap \overline{Y}|$ est le nombre d'attributs qui ne sont présents ni dans x ni dans y. Finalement, $e = |X \cup Y|$ représente le nombre d'attributs qui sont dans x ou dans y.
Plusieurs mesures de similarité ont été proposées sur ce genre de données, selon que l'on ne prend pas en considération le facteur d (attributs non présents dans les deux objets) et dans ce cas les mesures de similarité sont dites de type 1, ou dans le cas contraire, où les mesures de similarité sont dites de type 2 [LRB09b]. La Table (7) présente une chronologie des mesures de similarité qui ont été proposées.
Dans le cas de données numériques, chaque objet est représenté sous forme d'un vecteur d'attributs définis dans $\chi = \mathbf{R}^p$. Sur ce type de données, il n'est pas possible d'appliquer des mesures de similarité exploitant les quantités ensemblistes (a,b,c,d,e) présentées auparavant. Pour cela, la similarité entre deux objets x et y est quantifiée via une fonction SIM (9) qui est exprimée soit en relation à une expression de distance, soit par le produit scalaire de leurs vecteurs représentatifs

Mesure de similarité de type 1	Formule (SIM)
Jaccard (1908)	$\frac{a}{a+b+c}$
Kulczynski (1927)	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$
Dice (1945)	$\frac{2a}{2a+b+c}$
Sorensen (1948)	$\frac{4a}{4a+b+c}$
Ochiai (1957)	$\frac{a}{\sqrt{a+b+\sqrt{a+c}}}$
Anderberg (1973)	$\frac{8a}{8a+b+c}$
Sneath et Sokal (1973)	$\frac{a}{a+2(b+c)}$
Tversky (1977)	$\frac{a}{a+\alpha.b+\beta.c}$
Mesure de similarité de type 2	**Formule**
Russel et Rao (1940)	$\frac{a}{a+b+c+d}$
Yule et Kendall (1950)	$\frac{ad}{ad+bc}$
Sokal et Michener (1958)	$\frac{a+d}{a+b+c+d}$
Rogers et Tanimoto (1960)	$\frac{a+d}{a+2(b+c)+d}$
Sokal et Sneath (1963)	$\frac{a+d}{a+\frac{1}{2}(b+c)+d}$

TABLE 7 – Chronologie des mesures de similarité de type 1 (haut), et de type 2 (bas du tableau). Certaines mesures de similarité peuvent varier dans [-1,1], notamment pour les mesures de similarité du deuxième type. Ces mesures de similarité sont appliquées sur des données binaires [LRB09b].

Distance	Formule
Euclidienne	$d(x,y)=\sqrt{\Sigma_{i=1}^{p}(x_i-y_i)^2}$
Euclidienne pondérée	$d(x,y)=\sqrt{\Sigma_{i=1}^{p}\alpha_i(x_i-y_i)^2}$
Mahalanobis	$d(x_i,y_j)=\sqrt{(x_i-y_j)^t\Sigma^{-1}(x_i-y_j)}$, Σ : matrice de covariance
Minkowski	$d_\gamma(x,y)=(\Sigma_{i=1}^{p}\lvert x_i-y_i\rvert^\gamma)^{\frac{1}{\gamma}}$, γ : paramètre positif.
Manhattan	$d(x,y)=\Sigma_{i=1}^{p}\lvert x_i-y_i\rvert$ (Minkowski pour γ=1)

TABLE 8 – Différentes mesures de distance entre objets caractérisés par des attributs numériques [LRB09b].

[LRB09b] :

$$\begin{array}{rcl} SIM : \mathbf{R}^p \times \mathbf{R}^p & \longrightarrow & [0,1] \\ (x,y) & \longmapsto & SIM(x,y) \end{array}. \tag{9}$$

Les distances qui sont souvent utilisées sur des données numériques sont présentées dans la Table 8. A partir des valeurs de distances obtenues avec l'une des fonctions présentées dans ce tableau, la valeur de similarité entre deux objets x et y s'exprime avec une fonction f décroissante $SIM(x,y) = f(d(x,y))$. Cette fonction f est le plus souvent le complément à 1 de la fonction de distance d [LRB09b].
L'autre façon d'exprimer une similarité, se fait par le calcul du produit scalaire entre deux vecteurs. Le produit scalaire est une fonction qui est croissante quand les deux vecteurs d'objets sont similaires. Il dépend de l'angle qui sépare les deux vecteurs et de leurs normes : $< x, y >= ||x||.||y||.cos(x,y)$. On peut en déduire la valeur de la similarité qui se calcule par le

cosinus des deux vecteurs : $SIM(x,y) = \frac{<x,y>}{||x||.||y||}$.
Le choix d'une mesure de similarité dépend fortement de la nature des attributs. En effet, nous avons vu que si les attributs sont de nature binaires, les métriques présentées dans la Table 7 peuvent être appliquées. Dans le cas des attributs numériques multivalués, les mesures de similarité se calculent soit par une distance, soit par un produit scalaire.

4.3.2 Définition des mesures de similarité sémantique dans Gene Ontology

Dans le cas où les attributs qui décrivent les objets comparés sont structurés dans une hiérarchie de termes, ordonnés par des relations sémantiques, on fait souvent référence à la notion de *mesure de similarité sémantique*. Par exemple, dans le domaine de la recherche d'information, la comparaison entre des documents qui sont décrits par des termes de thésaurus ou d'une ontologie implique l'utilisation d'une mesure de similarité sémantique [LSSM08].
Plusieurs domaines ont vu l'application des mesures de similarité sémantique. Par exemple des auteurs comme *Jiang* [JC97] et *Resnick* [Res95] sont considérés comme les pionniers à avoir utilisé ces mesures sémantiques dans le domaine linguistique de la langue anglaise, en utilisant le vocabulaire contrôlé *WordNet* [Mil95].
Posons O une ontologie du domaine, telle que : O=$(C, \sqsubseteq_C, A)$, où C=$\{c_1, c_2, ..., c_n\}$ est un ensemble fini de concepts, A est l'ensemble des attributs des concepts, et $\sqsubseteq_C$ est une relation d'ordre partiel sur les concepts (réflexive, anti-symétrique). Un objet (x) peut être annoté par une ou plusieurs instances de concepts de l'ontologie O. Blanchard *et al.* ont présenté un modèle général de mesure de similarité sémantique appliqué à une hiérarchie [BHK08]. Selon ces auteurs ces mesures sémantiques se distinguent en deux catégories, d'un côté celles qui dépendent seulement des relations hiérarchiques entre les termes d'annotation [WP94], et d'un autre côté celles qui intègrent d'autres facteurs statistiques comme la fréquence des termes d'annotation et le contenu d'information [Lin98].
Dans le domaine biomédical, la notion de *similarité fonctionnelle* a été introduite pour décrire la similarité entre des produits de gènes qui se calcule par agrégation de la mesure de similarité sémantique entre leurs annotations issues de l'ontologie GO (O_{GO}). En effet, les biologistes ont souvent tendance à calculer la similarité entre des gènes pour diverses raisons, par exemple l'identification de corrélation entre l'expression des gènes et leurs fonctions biologiques (termes GO) [SSP+05, BW07].
Notons g_1 et g_2 deux produits de gènes annotés par deux listes de termes GO : g_1=$\{t_{1,1}, ..., t_{1,i}, ..., t_{1,n}\}$ et $g_2 = \{t_{2,1}, ..., t_{2,i}, ..., t_{2,m}\}$. Soit le terme LCA (Lowest Common Ancestor) un terme d'annotation qui est l'ancêtre commun le plus spécifique des deux termes d'annotation t_1 et t_2. La spécificité d'un terme est en relation avec sa profondeur par rapport à la racine dans le graphe de l'ontologie. Plus un terme est profond plus il est spécifique, et inversement pour la généralité du terme.
Comme nous l'avons déjà présenté, l'ontologie GO est organisée sous forme de rDAG, avec des relations hiérarchiques. Par conséquent, la similarité fonctionnelle entre deux produits de gènes $(SIM(g_1, g_2))$ peut être quantifiée à partir des mesures de similarité sémantique entre les termes qui les annotent : $SemSIM(t_i, t_j)$, $\forall\ t_i \in g_1$, $\forall\ t_j \in g_2$.
Dans un travail très récent, Pesquita *et al.* ont défini une mesure de similarité sémantique comme étant une fonction $SemSIM$, qui pour un couple de termes de l'ontologie GO $(t_i,t_j) \in O^2_{GO}$ retourne une valeur numérique qui reflète la proximité sémantique entre les deux termes [PFF+09].
Les mêmes auteurs distinguent la comparaison entre les termes de l'ontologie (instances de concept) et entre des ensembles de termes d'annotation, où chaque ensemble représente la liste

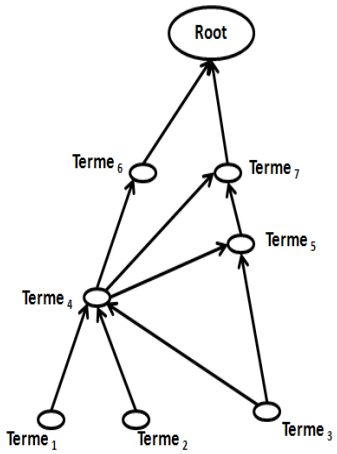


FIGURE 11 – Exemple d'un ensemble de termes d'annotations structurés dans le graphe de l'ontologie.

de termes qui annotent un objet (produit de gène par exemple).

4.3.2.1 Comparaison entre deux termes d'annotation

Dans le cas de la similarité sémantique entre deux termes d'annotation pris individuellement de l'ontologie GO, Pesquita *et al.* [PFF+09] distinguent de la même façon que Blanchard *et al.* [BHK08] deux catégories : la première représente les mesures de similarité sémantique basées sur les arcs du graphe de l'ontologie (*edge-based*) dont le principe repose sur le comptage des arcs séparant les deux termes d'annotation. La deuxième catégorie concerne les mesures qui sont basées sur l'information qui réside dans les nœuds du graphe de l'ontologie (*node-based*), et qui exploitent l'information que contient le nœud qui représente un terme d'annotation, ses parents et ses descendants.

Dans la première catégorie de mesures de similarité sémantique (*edge-based*), parmi les plus utilisées sont celles basées sur le calcul du *plus court chemin* (SPL : "Shortest Path-Length") entre les deux nœuds du graphe représentant les deux termes d'annotation comparés. Cet indicateur a été utilisé par *Rada et al.* [RMBB89] pour calculer la similarité sémantique entre les termes d'annotations de l'ontologie MeSH (Medical Subject Headings), et par *Nagar et* Al-Mubaid [NAM08b] sur des termes de l'ontologie GO.

Pour calculer la similarité sémantique entre deux termes d'annotation t_i et t_j avec la méthode de *Nagar* et d'*Al-mubaid* (dite ci-après méthode d'*Al-Mubaid*), le graphe de l'ontologie GO est parcouru afin de minimiser le nombre d'arcs qui séparent les termes, c'est à dire la longueur du chemin. Plusieurs algorithmes ont été proposés pour calculer le plus court chemin (SPL) dans un graphe [Joh73]. Le principe général est d'extraire pour deux termes (t_1, t_2), la liste de leurs parents communs. Ensuite, on calcule la longueur de tous les chemins possibles ayant comme extrémité les deux termes et passant par un parent partagé. Par conséquent le SPL est le plus petit de ces chemins.Par exemple, dans la Figure 11 la liste des parents partagés pour $Terme_1$ et $Terme_3$ sont $\{Terme_4, Terme_5, Terme_7, Root\}$ et la longueur des chemins passant par ces parents sont (2,3,4,5) respectivement. Le plus court chemin est donc $SPL = 2$. Pour en faire des valeurs obtenues avec cette mesure des valeurs de similarité plus qu'une distance, une fonction de transfert est utilisée [NAM08b].

L'un des inconvénient de cette mesure réside dans la non prise en considération de la profondeur

des termes ou de leurs parents partagés. En effet, dans la Figure 11, $SemSIM_{Al-Mubaid}(Terme_1, Terme_2)= 2$ à la même valeur que $SemSIM_{Al-Mubaid}(Terme_6, Terme_7)$. Pourtant, $Terme_1$ et $Terme_2$ sont plus spécifiques que ($Terme_6$ et $Terme_7$).
Concernant la deuxième catégorie des mesures de similarité sémantique terme à terme, c'est-à-dire celles qui sont basées sur l'information qui réside dans les nœuds du graphe de l'ontologie (*node-based*), ce sont probablement les plus citées dans les travaux qui concernent ce domaine. Les mesures de cette catégorie se basent sur la théorie de *Shannon* pour calculer le contenu en information (*IC : Information Content*) d'un terme d'annotation [SWS98]. Ainsi ayant deux termes d'annotation à comparer, le contenu en information est calculé pour les deux termes et pour leur parent partagé le plus spécifique (LCA) [Res95, Lin98]. Le contenu en information d'un terme est basé sur sa fréquence, ou sa probabilité d'occurrence (p) dans un corpus, *i.e.*, $IC(t_i)=-Log(p(t_i))$ [Res95, Flo04]. De ce fait, un terme avec une grande fréquence dans un corpus, dispose d'un contenu d'information faible et vice-versa. Intuitivement, les termes qui sont spécifiques c'est à dire situés profondément dans l'ontologie, ont un IC très élevé.
Resnik est considéré comme le premier à avoir introduit dans la définition de sa mesure de similarité sémantique le facteur IC calculé pour l'ancêtre commun le plus spécifique (LCA) de deux termes t_i et t_j [Res95]. Puisque ce facteur croît quand la profondeur augmente, le contenu en information du LCA est forcément le plus grand par rapport à ceux de tous les ancêtres des deux termes. La mesure de Resnik est définie par l'équation suivante :

$$SemSIM_{Resnik}(t_i, t_j) = IC(LCA(t_i, t_j)). \tag{10}$$

La formule (10) signifie que si le contenu en information du LCA des deux termes comparés augmente, cela implique une grande similarité sémantique entre les deux termes. Une des limitations de la mesure de *Resnik* réside dans le fait que tous les couples de termes qui ont un LCA identique, même s'ils ont des contenus d'information différents, seront identiquement similaires. De la même façon, par exemple, en se référant à la Figure 11, $SemSIM_{Resnik}(Terme_6, Terme_7)$ et $SemSIM_{Resnik}(Terme_6, Terme_5)$ sont équivalentes à $IC(Root)$. Pourtant la profondeur (spécificité) de $Terme_5$ est plus grande que celle de $Terme_7$, par conséquent $IC(Terme_5) > IC(Terme_7)$. Pour remédier à ce problème, *Jiang et Conrath* ont proposé une nouvelle mesure de similarité sémantique [JC97], adaptée de celle de *Resnik*, incluant le contenu en information des deux termes comparés :

$$SemSIM_{JC}(t_i, t_j) = IC(t_i) + IC(t_j) - 2 * IC(LCA(t_i, t_j)). \tag{11}$$

Puisque les valeurs prises par le contenu en information ne sont pas supérieurement bornées, *Lin* a introduit une normalisation sur la mesure de *Resnik* en divisant par la somme du contenu en information des deux termes comparés. La mesure sémantique de Lin [Lin98] est définie par la formule suivante :

$$SemSIM_{Lin}(t_i, t_j) = 2 * \frac{IC(LCA(t_i, t_j))}{IC(t_i) + IC(t_j)}. \tag{12}$$

Un inconvénient de la mesure de Lin, s'illustre dans le cas où $IC(Terme_i)$, $IC(Terme_j)$, $IC(LCA\ (Terme_i, Terme_j))$ sont proportionnels à $IC(Terme_{i'})$, $IC(Terme_{j'})$, $IC(LCA\ (Terme_{i'}, Terme_{j'}))$, alors $SemSIM_{Lin}(Terme_i, Terme_j) = SemSIM_{Lin}(Terme_{i'}, Terme_{j'})$, or $Terme_{i'}$, $Terme_{j'}$ pourraient être plus spécifiques que $Terme_i$ et $Terme_j$, et donc plus similaires sémantiquement. Pour faire face à ce problème, Schlicker *et al.* ont récemment apporté des modifications sur la mesure de *Lin* en introduisant un facteur de correction basé sur la probabilité d'occurrence du LCA des deux termes [SDRL06]. Cette mesure est définie par :

$$SemSIM_{Schlicker}(t_i, t_j) = SemSIM_{Lin}(t_i, t_j) * (1 - P(LCA(t_i, t_j)). \tag{13}$$

Ainsi, plus $P(LCA(t_i, t_j)$ est faible, c'est-à-dire plus le LCA de t_i et t_j est spécifique, plus la similarité sémantique entre ces deux termes est grande. Une des limitations des mesures basées sur le contenu en information (*node-based*), réside dans le fait qu'elles ne considèrent pas explicitement la profondeur des termes dans le graphe de l'ontologie, ni la distance qui les sépare de leurs LCA [PFF+09]. Des approches hybrides combinant à la fois le contenu en information et la profondeur des termes ont été proposées par la suite, comme c'est le cas dans les mesures de Wang *et al.* [WDP+07] et Othman *et al.* [ODI08]. La mesure de Wang introduit la notion de contribution sémantique d'un terme, où un poids est affecté à chaque arc reliant le terme à la racine du graphe par divers chemins. La contribution sémantique d'un terme est calculée comme le maximum de la somme des poids dans les différents chemins possibles jusqu'à la racine. Par conséquent la similarité sémantique selon la formule de Wang est la somme des contributions sémantiques de tous les ancêtres communs des deux termes divisée par la somme de la contribution sémantique de tous les ancêtres. Quant à la mesure d'Othman, les auteurs calculent la somme du contenu en information des termes qui séparent le premier terme de son LCA plus la somme des IC des termes qui se trouvent dans le chemin séparant le deuxième terme avec le même LCA.

4.3.2.2 Comparaison entre ensembles de termes d'annotation

Dans cette section, on considère qu'un produit de gène est annoté par un ensemble de termes. Par conséquent, pour comparer deux produits de gènes, *la similarité fonctionnelle* se ramène à la comparaison de deux ensembles de termes d'annotations. Le problème posé concerne donc la façon d'agréger les similarités sémantiques entre les termes. Pour cela, Pesquita *et al.* dans leur revue distinguent deux catégories de mesures de similarité entre ensemble de termes d'annotation. La première catégorie regroupe les méthodes qui calculent la similarité sémantique entre tous les paires possibles de termes d'annotation des deux ensembles : ce sont les méthodes dites à paires de termes (*pairwise* methods). La deuxième catégorie quant à elle concerne les méthodes qui considèrent les ensembles de termes d'annotation comme un seul ensemble mathématique, un vecteur ou un graphe (*groupwise* methods).
Considérons les deux produits de gènes g_1 et g_2 avec leurs listes de termes d'annotations : g_1=$\{t_{1,1}, ..., t_{1,i}, ..., t_{1,n}\}$ et $g_2 = \{t_{2,1}, ..., t_{2,i}, ..., t_{2,m}\}$ respectivement. Calculer la similarité sémantique entre toutes les paires possibles des termes d'annotation signifie la génération d'une matrice de similarité sémantique $MatSim$. Il existe trois approches d'agrégation des valeurs

	j=1	j=2	...	j=m	MaxLigne
i=1		...	⋮	...	
i=2		⋮		...	
...		⋮	$MatSim[x, y] = SemSim(t_{1,x}, t_{2,y})$		$MaxLigne[x]=$ $Max(MatSim[k, x])\forall k\in[1, m]$
...		⋮		...	
i=n		⋮		...	
MaxColonne	...		$MaxColonne[y]= Max(MatSim[h, y])\forall h\in[1, n]$	...	$BMA = \frac{1}{2}(\frac{\sum_{i=1}^{n}}{n} Maxligne[i]$ $+\frac{\sum_{j=1}^{m}}{m} MaxColonne[j])$

TABLE 9 – Matrice de similarité sémantique $MatSim$ entre les paires de termes qui annotent deux produits de gènes. Chaque cellule $MatSim[x, y]$ représente la similarité sémantique entre le terme $t_{1,x}$ qui annote g_1, et le terme $t_{2,y}$ qui annote g_2.

calculées de $SemSIM$ pour obtenir une seule valeur $SIM(g_1, g_2)$: l'utilisation de la moyenne,

l'extraction du Min-Max, et le meilleur score moyen (BMA : Best Matching Average).
L'agrégation par moyenne signifie que la similarité sémantique est calculée entre toutes les paires de termes, et qu'une seule valeur moyenne est obtenue à la fin qui représente la similarité entre les deux produits de gènes comparés.

$$SIM(g_1, g_2) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} SemSIM(t_{1,i}, t_{2,j})}{n * m}. \tag{14}$$

L'agrégation par Min-Max, suggère de calculer soit le minimum, soit le maximum de toutes les valeurs de similarité sémantique calculées.

$$SIM(g_1, g_2) = Max|Min(SemSIM(t_{1,i}, t_{2,j})), \quad \forall i \in [1, n], \forall j \in [1, m]. \tag{15}$$

Finalement, pour l'agrégation par meilleur score moyen (BMA), comme dans la Table 9, on calcule les maximums dans les lignes et les colonnes de la matrice de similarité, puis on fait la moyenne de ces valeurs maximales.
Concernant l'agrégation par la moyenne, *Lord et al.* ont été les premiers à l'avoir utilisée pour une similarité sémantique sur l'ontologie GO [LSBG03]. En effet, la similarité sémantique entre des protéines est calculée en utilisant la mesure de *Resnik* entre toutes les paires de termes GO. Ensuite une valeur moyenne est calculée sur ces valeurs. La mesure de *Lord* est basée sur l'utilisation du contenu en information, elle est donc (*Average Pairwise Node-based*). La mesure de similarité fonctionnelle de Lord entre deux produits de gènes est définie par la formule suivante :

$$SIM_{Lord}(g_1, g_2) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} SemSIM_{Resnik}(t_{1,i}, t_{2,j})}{n * m}. \tag{16}$$

Nagar et Al-Mubaid *et al.* [NAM08b], [NAM08a] dans le même contexte ont utilisé la moyenne de leur similarité sémantique entre les termes d'annotations. Pour rappel, leur similarité sémantique est basée sur le calcul du plus court chemin entre deux termes, par conséquent elle est (*Average Pairwise Edge-based*). Dans la formule (17), une fonction de transfert f est appliquée sur la mesure d'Al-mubaid pour transformer la valeur calculée en une valeur de similarité. Cette fonction est définie de telle sorte que lorsque la moyenne des tailles des chemins séparant les termes d'annotation est grande, alors la similarité décroît. C'est par exemple $f{=}e^{-x}$.

$$SIM_{Al-Mubaid}(g_1, g_2) = f(\frac{\sum_{i=1}^{n} \sum_{j=1}^{m} SemSIM_{Al-Mubaid}(t_{1,i}, t_{2,j})}{n * m}), \tag{17}$$

D'autres mesures de similarité fonctionnelle utilisent l'agrégation par calcul du Max-Min, c'est à dire que la similarité fonctionnelle entre deux produits de gènes est équivalente à une valeur extrême (Min ou Max) de toutes les valeurs de similarité sémantique entre les termes d'annotations [SFSZ05, SSZ04].

$$SIM_{i,j}(g_1, g_2) = Max(SemSIM(t_{1,i}, t_{2,j})), . \tag{18}$$

Schlicker et al [SDRL06], utilisent un score BMA sur les valeurs de similarité sémantique calculées avec leur mesure définie dans la formule (13). Ayant deux produits de gènes g_1 et g_2, chaque valeur de la matrice de similarité $MatSim[x, y]$ est égale à $SemSIM_{Schlicker}(t_{1,x}, t_{2,y})$, pour avoir la similarité fonctionnelle $SIM_{Schlicker}(g_1, g_2){=}\ BMA$ (voir Table 9).
Dans la catégorie de mesures qui considèrent les annotations comme un groupe de termes (*group-wise* methods), on peut trouver des mesures de similarité qui sont sémantiques ou non, selon notamment que le groupe de termes considéré est ou non enrichi en ajoutant soit tous les ancêtres, soit tous les descendants des termes d'annotation.

Les auteurs comme *Lee et al.* [LHS+04], ont proposé l'approche la plus simple dans cette catégorie, qui est basée sur le calcul de l'intersection des deux ensembles des termes d'annotation qui annotent les deux produits de gènes comparés : $SIM(g_1, g_2)$=$g_1 \cap g_2$. Cette approche est clairement non sémantique puisqu'elle ne considère pas les relations sémantiques entre les annotations [BHK08]. *Martin et al.* dans une solution très similaire, ont utilisé la méthode de *Jaccard* pour définir une similarité, et la méthode de *Czekanowski-Dice* pour définir une distance dans leur outil GOToolBox [MBR+04] (et Table 7). La similarité sémantique est utilisée ici en ce sens que chaque liste de terme est étendue par la liste de tous les ancêtres des termes annotant les gènes. Notons g_i^* l'ensemble ainsi étendu des termes d'annotation pour le gène g_i. Ainsi la similarité et la distance fonctionnelle entre deux produits de gènes, utilisées dans cet outil se calculent par ces formules :

$$SIM_{UI}(g_1^*, g_2^*) = \frac{|g_1^* \cap g_2^*|}{|g_1^* \cup g_2^*|}, \quad Dist(g_1^*, g_2^*) = \frac{|g_1^* \Delta g_2^*|}{|g_1^* \cap g_2^*| + |g_1^* \cup g_2^*|}. \tag{19}$$

avec $g_1 \Delta g_2$ représentant les termes qui sont dans la différence des deux ensembles.
Une autre mesure de similarité ensembliste a été proposée par Guo *et al.*, en calculant l'union-intersection sur les chemins partagés entre les annotations. Cette similarité qui s'appelle *SimUI* peut être utilisée pour mesurer la similarité entre deux graphes. Les auteurs l'ont testée pour constituer des réseaux métaboliques de gènes [GLS+06].
Récemment, *Pesquita et al* ont apporté une amélioration sur la mesure *SimUI* en pondérant les termes d'annotation par leur contenu en information [PFB+08]. La similarité proposée s'appelle *SimGIC (Graph Information Content)* ou Jaccard pondéré.

$$SIM_{GIC}(g_1, g_2) = \frac{\sum_{t \in g_1 \cap g_2} IC(t)}{\sum_{t \in g_1 \cup g_2} IC(t)}. \tag{20}$$

Une autre sous-catégorie de mesures de similarité ensemblistes opte pour une modélisation vectorielle inspirée des méthodes de la recherche documentaire [BAB05, SM83, Pol04]. Les gènes sont représentés dans un modèle d'espace vectoriel (VSM : Vector Space Model) sous forme de vecteurs $\vec{g}$ où chaque composante correspond à un terme d'annotation et indique si ce terme annote ou non le gène. L'espace est de dimension k, représentant les k termes d'annotation et les vecteurs de bases sont $\vec{e_i}$ [GAM+03, TP09]. Ce modèle est souvent utilisé en recherche d'information pour calculer la similarité entre des documents textuels. La similarité est calculée comme la valeur du cosinus entre les deux gènes. Ce cosinus suggère de calculer le produit scalaire entre les termes d'annotations. *Chabalier et al* ont utilisé cette formulation pour proposer leur mesure de similarité vectorielle, en pondérant chaque terme t_i de la dimension (i) par un poids w_i, qui représente le facteur IDF (Inverse Document Frequency) de ce terme par rapport à un corpus [GAM+03, CMB07]. La valeur IDF d'un terme est le nombre total de produit de gènes d'un corpus divisé par le nombre de gènes de ce corpus annotés par ce terme, dans une échelle logarithmique.

$$SIM_{Vect}(g_1, g_2) = \frac{\vec{g_1}.\vec{g_2}}{||\vec{g_1}||.||\vec{g_2}||}, \tag{21}$$

où $\vec{g_1}.\vec{g_2}$=$\sum_{i=1} kw_{1,i}.w_{2,i}$ (avec $w_{1,i}$=$TF_{1,i}.IDF_{1,i}$) pour tout terme t_i présent à la fois dans les annotations de g_1 et g_2.
L'inconvénient de la mesure de formulation vectorielle choisie par *Chabalier et al.* réside dans le fait qu'ils considèrent que l'espace vectoriel représentant les annotations est orthogonal, ce qui veut dire que le produit scalaire $t_{1,i}.t_{2,j} = 0$ pour $i \neq j$. Ceci implique qu'il n'y a aucune similarité sémantique entre $t_{1,i}$ et $t_{2,j}$, or ce n'est pas le cas si $t_{1,i}$ et $t_{2,j}$ proviennent d'un vocabulaire

structuré, voire d'une ontologie.

4.3.3 Discussion

La majorité de ces mesures de similarité sémantique utilisent les annotations GO comme connaissance du domaine. Ainsi, l'intérêt majeur de ces mesures de similarité consiste, dans le cas des données d'expression, à compléter les approches d'enrichissement statistique des annotations. Entre autres, les mesures de similarité sont utilisées pour regrouper les gènes et leurs annotations fonctionnelles dans des clusters distincts. Nous avons vu que pour comparer des termes d'annotations dans une ontologie ou dans un vocabulaire structuré, on utilise la notion de mesure de similarité sémantique. À un autre niveau, on utilise la notion de similarité fonctionnelle lorsque l'on calcule la similarité entre produits de gènes annotés par des termes d'annotations fonctionnelle. Les mesures de similarité sémantique diffèrent selon que l'on calcule la similarité sur chaque paire de termes (*pairwise*), ou que l'on considère les annotations comme un groupe (*groupwise*). Diverses informations sont utilisées pour mesurer la similarité sémantique terme-terme, par exemple : le contenu en information (*node-based*) ou la distance séparant les termes dans le graphe de l'ontologie (*edge-based*). L'agrégation des mesures de similarité sémantique en mesures fonctionnelles peut être calculée par la moyenne (*average*), le *Min-Max*, ou par calcul du meilleur score moyen (*BMA*).
Il faut noter qu'une mesure de similarité sémantique floue, a été proposée par *Popescu et al.* [PKM06]. Les auteurs ont utilisé l'intégrale de *Choquet* [GL05] pour définir leur mesure.
En plus des annotations GO, *Huang et al.* dans la plateforme DAVID ont pris en considération d'autres sources de connaissance [SHT+07], pour calculer la similarité fonctionnelle entre produits de gènes en utilisant la statistique Kappa [HST+07, WC05, DSH+03]. L'outil proposé est considéré comme la référence chez les biologistes pour réaliser la classification fonctionnelle de gènes et l'interprétation des données d'expression. Pour cette raison, nous allons consacrer la suite de ce chapitre à présenter et introduire la plateforme DAVID et le principe de base de l'outil de classification fonctionnelle. En effet, cet outil sera considéré comme la référence pour évaluer notre travail concernant la nouvelle mesure de similarité sémantique que nous avons proposée, et les résultats de la classification fonctionnelle.
Enfin, notons que la majorité des mesures de similarité sémantique présentées, sont disponibles dans le package *GOstats* et *csbl* sous R Bioconductor [GOv, Ova, OLH08].

4.4 Clustering fonctionnel de gènes et présentation de l'outil DAVID

Dans la dernière décennie de plus en plus de ressources biologiques ont vu le jour, pour servir dans la recherche biomédicale. Ces ressources distribuées, sont souvent hétérogènes et redondantes. Les biologistes avaient alors des difficultés pour mener à bien leurs expériences en utilisant plusieurs sources d'annotations. L'intérêt d'une base de connaissances centralisée contenant les annotations fonctionnelles des gènes s'avérait alors indispensable. Pour ces diverses raisons, la base de données DAVID (Database for Annotation Visualisation and Integrated Discovery) a été proposée en intégrant plusieurs sources de connaissances dans une seule plateforme [SHT+07, DSH+03]. Dans cette base, des millions d'identifiants de gènes et de protéines, qui sont obtenus de diverses bases de données, sont annotés par plusieurs milliers de termes fonctionnels provenant de 14 bases de données différentes (GO database, KEGG Pathways, BioCarta Pathways...) comme illustré dans la Figure 12.
De plus, la plateforme DAVID intègre deux fonctionnalités très puissantes : *DAVID Gene Func-*

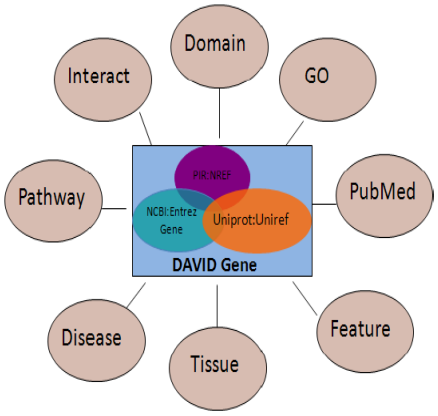


FIGURE 12 – Extrait de la base de connaissance de l'outil DAVID (Database for Annotation Visualisation and Integrated Discovery). Plusieurs sources sont utilisées pour annoter les gènes et les protéines [SHT+07].

tional Classification Tool pour le clustering des gènes et *DAVID Functional Annotation Clustering Tool* pour le clustering des termes d'annotations [HST+07]. La plateforme est disponible en ligne à l'adresse : http ://david.abcc.ncifcrf.gov/.
L'outil DAVID est maintenant considéré comme un des programmes d'annotation fonctionnelle les plus populaires auprès des biologistes et s'appuie sur une annotation fonctionnelle des gènes particulièrement riche. À chaque produit de gène est associé un *profil fonctionnel*, c'est à dire un vecteur d'annotation fonctionnelle, comme présenté sur la Figure 13 (a). Comme toute méthode de clustering fonctionnel [OLH08, SFSZ05, AS04], DAVID construit une matrice de similarité calculée entre toutes les paires de gènes, mais l'outil utilise la statistique *Kappa* [Coh60].
En partant d'une matrice d'annotations d'une liste de gènes comme dans la Figure 13 (a), on construit l'hypothèse qui suppose que si deux gènes ont des profils d'annotations similaires alors ils devraient être fonctionnellement similaires. La méthode permet entre autres d'identifier les groupes de gènes partageant une grande quantité d'annotations fonctionnelles. Le nombre de colonnes de cette matrice est égal au nombre total de termes d'annotations dans toute la base de connaissance utilisée par DAVID, de diverses sources : GO Biological Process, GO Molecular Function, GO Cellular Component, KEGG Pathways, BioCarta Pathways, Swiss-Prot Keywords, BBID Pathways, SMART Domains, NIH Genetic Association DB, UniProt Sequence Features, COG/KOG Ontology, NCBI OMIM, InterPro Domains, et des PIR SuperFamily Names.
Par la suite, une statistique Kappa est appliquée à chaque couple de gènes pour mesurer le degré de co-occurrence des annotations de ces deux gènes par rapport au hasard, comme illustré dans la Figure 13 (b). La valeur O_{ab} représente la co-occurrence (concordance) observée dans la matrice d'annotation, c'est à dire le nombre de fois où les deux gènes a et b sont annotés par un même terme ou pas. La valeur A_{ab} représente le nombre de fois où les deux gènes sont concordants par chance. En utilisant ces deux valeurs, la valeur $Kappa$ est calculée pour représenter le degré de similarité entre les deux gènes :

$$K_{a,b} = \frac{O_{ab} - A_{ab}}{1 - A_{ab}}. \tag{22}$$

On peut en déduire que $K_{a,b} = 1$ si et seulement si les deux gènes sont en parfaite concordance. Après la génération de la matrice de similarité Kappa, il est possible de grouper les gènes qui présentent des valeurs Kappa supérieures à un certain seuil, dans un même cluster fonctionnel.

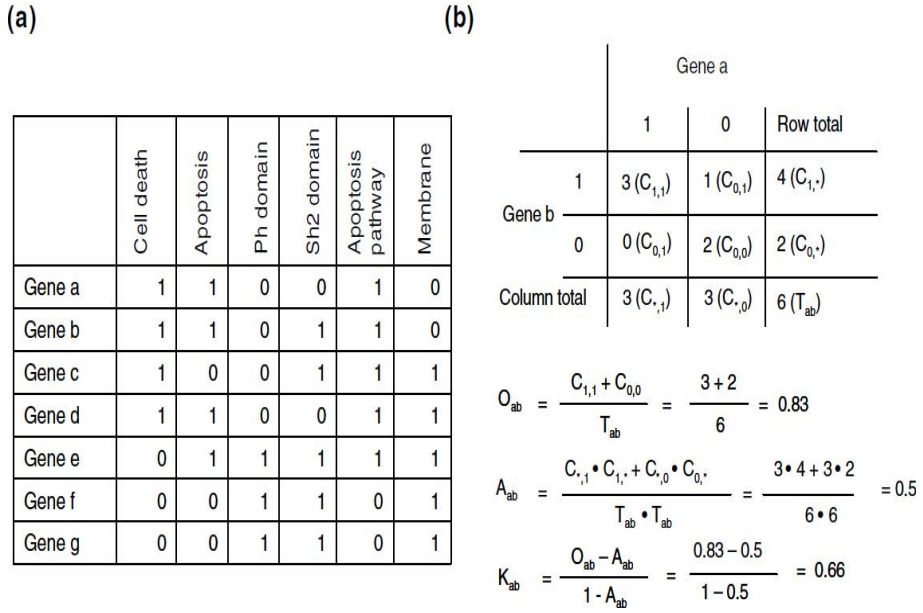
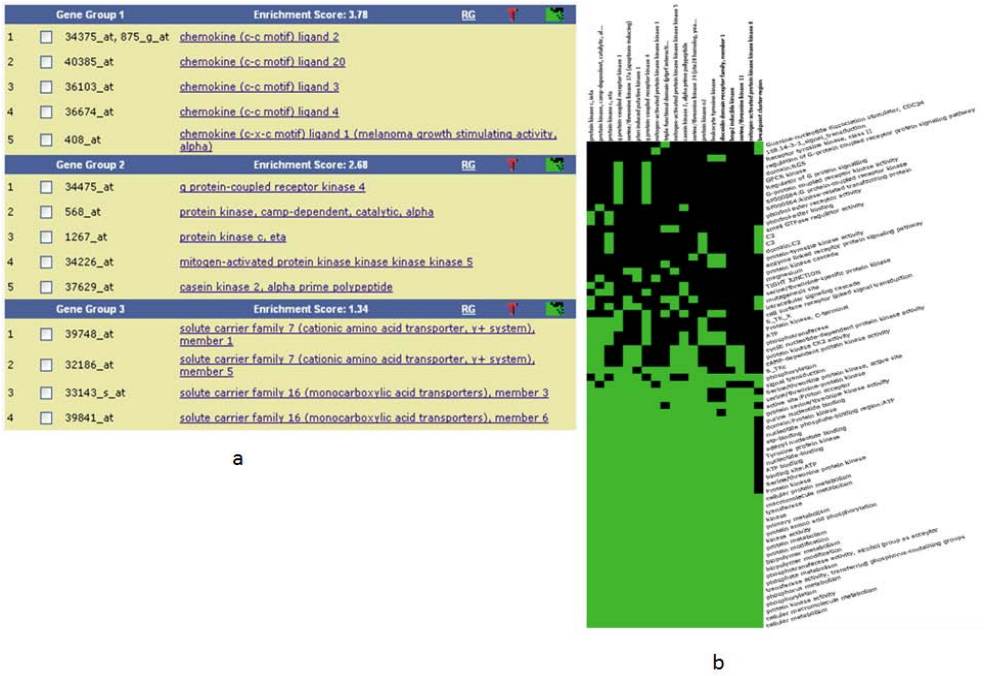


FIGURE 13 – Principe de base du fonctionnement de l'outil de DAVID dédié à la classification de gènes. (a) Pour chaque gène, un profil fonctionnel est constitué en marquant par 1 selon qu'un terme annote le gène et 0 dans le cas contraire. La matrice d'annotation est binaire. (b) Un exemple de table de contingence de deux gènes a et b, en utilisant la statistique Kappa qui calcule la concordance des annotations entre les deux gènes. La valeur Kappa calculée entre les deux gènes reflète leur degré de similarité. Ayant une liste de gènes, les valeurs Kappa peuvent être calculées pour chaque paire de gènes. La matrice de similarité Kappa est ensuite utilisée pour réaliser un clustering de gènes (Gene Functional Classification). Une rotation de la matrice d'annotations permet de calculer les similarités Kappa entre les termes d'annotations, et par conséquence leur utilisation pour le clustering de termes (Functional Annotation Clustering) [HST+07].



	Gene Group 1	Enrichment Score: 3.78	RG
1	34375_at, 875_g_at	chemokine (c-c motif) ligand 2	
2	40385_at	chemokine (c-c motif) ligand 20	
3	36103_at	chemokine (c-c motif) ligand 3	
4	36674_at	chemokine (c-c motif) ligand 4	
5	408_at	chemokine (c-x-c motif) ligand 1 (melanoma growth stimulating activity, alpha)	
	Gene Group 2	**Enrichment Score: 2.68**	**RG**
1	34475_at	g protein-coupled receptor kinase 4	
2	568_at	protein kinase, camp-dependent, catalytic, alpha	
3	1267_at	protein kinase c, eta	
4	34226_at	mitogen-activated protein kinase kinase kinase kinase 5	
5	37629_at	casein kinase 2, alpha prime polypeptide	
	Gene Group 3	**Enrichment Score: 1.34**	**RG**
1	39748_at	solute carrier family 7 (cationic amino acid transporter, y+ system), member 1	
2	32186_at	solute carrier family 7 (cationic amino acid transporter, y+ system), member 5	
3	33143_s_at	solute carrier family 16 (monocarboxylic acid transporters), member 3	
4	39841_at	solute carrier family 16 (monocarboxylic acid transporters), member 6	

a

b

FIGURE 14 – Exemple d'affichage des résultats du clustering fonctionnel de gènes avec l'outil DAVID. (a) : un affichage au format texte, dans chaque groupe de gènes on a la liste de gènes regroupés et les fonctions les plus enrichies. (b) : une matrice thermique (Heatmap), où en colonne on a la liste des gènes, les lignes représentent les annotations concernées. La couleur d'une cellule en vert représente le fait que le terme en question annote le gène de la colonne considéré, tandis que la couleur est noire quand il n'y a aucune relation entre le terme et le gène [HST+07].

Un algorithme de K-means flou (FCM) a été utilisé dans l'outil de classification DAVID afin de permettre à un gène d'appartenir à plusieurs clusters fonctionnels à la fois.
Les résultats de la classification de gènes peuvent être visualisés soit sous forme textuelle comme illustré dans la Figure 14 (a), soit sous forme de Heatmap (partie (b) de la même figure).
Bien que l'outil DAVID intègre une base de connaissance très riche, et utilise une mesure de similarité sophistiquée, cependant il représente cependant l'inconvénient majeur de ne pas tenir compte des relations sémantiques qui pourraient exister entre les annotations. En effet, certains termes d'annotations utilisés dans l'outil sont issus d'un vocabulaire structuré comme l'ontologie GO. L'outil suit un modèle d'annotation binaire, c'est à dire dans le cas où le gène n'est pas annoté par un terme d'annotation (x) alors dans la cellule adéquate de la matrice d'annotations (Figure (13)), on marquera zéro. Or, si le gène n'est pas annoté par ce terme (x), il pourrait être annoté par un terme sémantiquement similaire à ce terme, ce qui permettrait de réduire le nombre de zéros dans la matrice d'annotation.

5 Problématiques traitées dans la thèse et plan du mémoire

Au terme de cet état de l'art, je voudrais dégager les problématiques traitées dans cette thèse. Nous avons vu que l'interprétation de données transcriptomiques repose sur deux types d'analyse : la construction de profils d'expression des gènes d'une part, l'étude des annotations fonctionnelles, généralement issues de l'ontologie GO (Gene Ontology) des gènes d'autre part, le

deuxième type d'analyse servant à donner un sens au premier. J'ai voulu contribuer à l'efficacité de ce processus d'interprétation à quatre niveaux :

1. L'utilisation de mesures de similarité pour la classification fonctionnelle de gènes signatures n'était pas effective jusqu'à présent car la plupart des mesures de similarité sémantique n'étaient pas bien adaptées, en particulier elles ne respectaient pas la propriété de maximalité selon laquelle la similarité d'un objet avec lui même est maximale (égale à 1). J'ai donc défini une nouvelle mesure de similarité sémantique et fonctionnelle (IntelliGO), robuste et capable de considérer différemment les annotations fonctionnelles selon les différents aspects de GO, selon les espèces et aussi selon les codes d'évidence (chapitre 2).
2. La validation de la classification fonctionnelle grâce à des jeux de données constitués d'ensembles de référence m'a conduit à développer une méthode d'analyse de recouvrement entre clusters fonctionnels et ensembles de référence (chapitre 3).
3. Cette méthode d'analyse de recouvrement peut alors être appliquée au recouvrement entre clusters fonctionnels et profils d'expression, notamment dans le cas d'un jeu de données relatif au cancer colo-rectal. Pour cela, la notion de profil d'expression est revisitée avec l'introduction des profils flous d'expression différentielle (chapitre 4).
4. La mesure de similarité sémantique (IntelliGO) est ensuite généralisée à tout vocabulaire structuré en rDAG. La possibilité de réaliser alors un clustering sémantique des termes de l'ontologie a été utilisée sur l'ontologie Meddra (chapitre 5).

Enfin, je terminerai le manuscrit avec un bilan général et des perspectives (chapitre 6).

Chapitre 2

Une nouvelle mesure de similarité sémantique : IntelliGO

Sommaire

1 Présentation de la nouvelle mesure de similarité IntelliGO

1.1 Introduction

Les mesures de similarité sémantique présentées utilisent soit le contenu en information, soit la distance séparant les termes d'annotation de façon séparée, mis à part quelques mesures de similarité hybrides qui ont été discutées.
L'ontologie GO est structurée en trois aspects : BP (Biological Process), CC (Cellular Component), MF (Molecular Function). Les biologistes ont souvent tendance à interpréter les données biologiques en regardant un seul aspect à la fois. Les aspects sont orthogonaux, donc mesurer une similarité sémantique entre des annotations issues de deux aspects différents n'a pas vraiment de sens. De même il est connu que les gènes ne sont pas annotés de la même manière d'une espèce à l'autre. Il faut donc être prudent lorsque l'on interprète des similarités fonctionnelles entre gènes provenant d'espèces différentes. Malgré cela, parmi les mesures proposées, très peu permettent de spécifier l'aspect de GO considéré, ni une espèce précise.
Nous avons vu dans la section (3.2.2) du chapitre précédent, que les annotations GO sont qualifiées avec les codes d'évidence. Dans un calcul de similarité fonctionnelle entre deux gènes,

considérer de la même façon un terme attribué par un curateur et un autre de façon électronique, serait comme considérer que les deux termes sont attribués de la même façon. Or ce n'est pas le cas. À notre connaissance, il n'y a pas une mesure dans l'état de l'art qui permet de prendre en considération de façon différentielle les codes d'évidence.
Pour toutes ces raisons j'ai décidé de proposer une similarité sémantique et fonctionnelle, notée *IntelliGO* et qui tient compte à la fois de la diversité de l'ontologie GO en terme de ses trois aspects orthogonaux, de la variabilité des codes d'évidence et de la possibilité de spécifier une espèce donnée [BSTP+10].

1.2 L'espace vectoriel généralisé de la mesure IntelliGO

1.2.1 Le schéma de pondération des vecteurs d'annotation

Dans le cas des mesures de similarité sémantique qui suivent un modèle vectoriel, un espace vectoriel est utilisé pour représenter chaque gène sous forme de vecteur $\vec{g}$ où chaque dimension $\vec{e_i}$ de ce vecteur représente une annotation [GAM+03, CMB07]. Le modèle vectoriel a été utilisé depuis longtemps dans le domaine de la recherche d'information [SM83],[Pol04],[BAB05]. Dans notre contexte, les documents et les termes sont remplacés par des gènes et des termes d'annotations respectivement. Un vecteur-gène est donc défini comme suit :

$$\vec{g} = \sum_i \alpha_i * \vec{e_i}, \tag{1}$$

Où les vecteurs $\{\vec{e_1},\vec{e_2},\vec{e_3},...,\vec{e_n}\}$ constituent une famille génératrice qui veut dire que tout vecteur $\vec{g}$ peut être représenté comme combinaison linéaire des vecteurs $\{\vec{e_1},\vec{e_2},\vec{e_3},...,\vec{e_n}\}$. Le vecteur $\vec{e_i}$ correspond au terme t_i du vocabulaire d'annotation. La valeur α_i est un coefficient de pondération sur ce terme.
Le premier point d'innovation de la mesure *IntelliGO* réside dans son schéma de pondération qui inclut à la fois le contenu en information et l'origine des annotations fonctionnelle. Le coefficient α_i affecté à chaque composante du vecteur d'annotation (terme GO) est composé de deux valeurs analogues aux facteurs *tf et idf* utilisés dans la recherche d'information [BCG+05]. D'une part, un poids $w(g, t_i)$ est assigné à un code d'évidence (EC) qui reflète la qualité de l'annotation entre le gène g et le terme GO t_i. D'autre part, nous introduisons le facteur IAF (*Inverse Annotation Frequency*) qui pour un corpus d'annotation de gènes, reflète le ratio entre le nombre total de gènes G_{Tot} et le nombre de gènes G_{t_i} annotés par le terme t_i. La valeur du coefficient IAF d'un terme GO t_i est calculée avec la formule suivante

$$IAF(t_i) = log\frac{G_{Tot}}{G_{t_i}}. \tag{2}$$

On peut facilement en déduire que cette définition est similaire à ce qui a été introduit concernant le contenu en information des termes GO par rapport à un corpus. Par conséquent, on peut facilement vérifier que les termes GO qui sont très fréquents dans un corpus, auront un IAF faible, et les termes qui sont rares dans un corpus auront un IAF élevé.
En résumé, le coefficient de pondération α_i est défini comme :

$$\alpha_i = w(g, t_i) * IAF(t_i). \tag{3}$$

1.2.2 Définition de la similarité sémantique IntelliGO entre deux termes d'annotation

Le second point d'innovation dans l'espace vectoriel de la mesure *IntelliGO*, concerne les vecteurs générateurs. Dans un espace vectoriel classique, les bases sont orthonormées, i.e. les vecteurs de la base sont normés et mutuellement orthogonaux. Ceci veut dire que chaque dimension de l'espace vectoriel (dans notre cas chaque terme d'annotation) est indépendante de toutes les autres. Or dans le cas des annotations de gènes, cette hypothèse n'est pas vérifiée, puisque les termes GO sont sémantiquement reliés dans le rDAG de l'ontologie GO. Par conséquent, dans l'espace vectoriel de la mesure de similarité sémantique *IntelliGO* les vecteurs générateurs ne sont pas considérés comme mutuellement orthogonaux dans un aspect donné (BP, MC ou CC). Une situation similaire a été illustrée dans les travaux de Ganesan *et al.* [GGMW03] dans le contexte de la recherche d'information à partir de documents annotés par un vocabulaire hiérarchique structuré sous forme de plusieurs arbres et connue sous le nom de MeSH (Medical Subject Headings) [Lip].
Ayant deux termes d'annotation t_i et t_j, représentés par leurs vecteurs de base $\vec{e_i}$ et $\vec{e_j}$ respectivement, la valeur du produit scalaire entre ces deux vecteurs de base est donnée par la formule suivante

$$\vec{e_i} * \vec{e_j} = 2 * \frac{Depth(LCA(t_i, t_j))}{Depth(t_i) + Depth(t_j)}. \quad (4)$$

Les valeurs des produits $\vec{e_i} * \vec{e_j}$ sont ensuite utilisées pour calculer le cosinus généralisé (GCSM : Generalized Cosine-Similarity Measure) entre deux gènes (voir section suivante). L'application de cette mesure dans le cas d'un rDAG n'est pas triviale. En effet, comme nous l'avons mentionné, dans un rDAG il existe plus d'un chemin qui peut relier un terme à la racine *Root*. Ceci aura des conséquences sur la formule (4), parce qu'un couple de termes d'annotation pourra avoir plus d'un LCA, et deuxièmement la profondeur d'un terme dans le rDAG n'est pas unique puisque celle-ci dépend du chemin qui relie le terme à la racine du rDAG. Pour ces raisons, nous avons adapté la formule (4) de Ganesan pour qu'elle soit applicable à un rDAG. Ayant observé que dans un arbre on a l'égalité $Depth(t_i) + Depth(t_j) = 2 * Depth(LCA(t_i, t_j)) + SPL(t_i, t_j)$, nous avons cherché à remplacer ainsi dans le cas d'un rDAG l'expression $Depth(t_i) + Depth(t_j)$ par une expression qui implique seulement $Depth(LCA)$ et SPL. La démonstration suivante a été écrite dans un formalisme inspiré de celui de Couto *et al.* [CSC07] pour représenter les relations entre termes dans l'ontologie GO.
Le vocabulaire contrôlé de l'ontologie GO peut être défini par un triplet $\gamma=(T, \Xi, R)$, où T est l'ensemble des termes d'annotations, Ξ est l'ensemble représentant les deux relations hiérarchiques principales $\Xi=\{$*is-a, part-of*$\}$.

Le troisième élément R contient un ensemble de triplets $\tau=(t, t', \xi)$, où $t, t' \in T$, $\xi \in \Xi$ et $t \xi t'$. Notons que ξ est une relation orientée de type enfant-ascendant, et que $\forall\ \tau \in R$, la relation ξ entre t et t' est parfois de type *is-a* ou *part-of*. Dans le vocabulaire γ, le terme *Root* représente le nœud du plus haut niveau dans le graphe de l'ontologie GO. En effet, *Root* est le parent direct des trois nœuds qui représentent les trois aspects de l'ontologie, *i.e.* $Biological\ Process$, $Cellular\ Component$, and $Molecular\ Function$.
Le nœud *Root* n'a pas d'ancêtre dans l'ontologie, donc il n'y a pas dans la collection R un triplet où t=*Root*. Tous les termes GO de T sont en relation avec le nœud racine de l'ontologie via un des trois nœuds des trois aspects.
Soit $Parents$ une fonction qui pour chaque terme t retourne une liste de parents directs de ce terme :

$$Parents : T \longrightarrow \mathcal{P}(T),$$

$$Parents(t) = \{t' \in T \ \mid \ \exists \xi \in \Xi, \exists \tau \in R, \ \tau = (t, t', \xi)\}, \tag{5}$$

Où $\mathcal{P}(T)$ représente l'ensemble des sous ensembles de T. Notons que $Parents(Root)$=$\emptyset$. La fonction $Parents$ est utilisée pour définir une autre fonction $RootPath$ comme étant l'ensemble des chemins directs allant du terme $Root$ jusqu'au terme t (voir Figure 11, pour une illustration) :

$$RootPath : T \longrightarrow \mathcal{P}(\mathcal{P}(T)),$$

$$RootPath(t) = \begin{cases} \{\{Root\}\} & si\ t = Root \\ \{\{t_1, ..., t_n\} | & \\ (t_1 = Root) \wedge & \\ (t_n = t) \wedge & \\ (t_{n-1} \in Parents(t_n)) & \\ \wedge (\{t_1, ..., t_{n-1}\} \in & \\ RootPath(t_{n-1}))\} & sinon \end{cases} . \tag{6}$$

Ainsi, chaque chemin dans le graphe de l'ontologie qui est entre le terme $Root$ et le terme t est un ensemble de termes ordonnés par leurs indice i $\Phi \in RootPath(t)$. La longueur d'un chemin séparant un terme t du terme racine $Root$ est définie dans le graphe comme le nombre d'arcs qui connectent les nœuds du chemin Φ. Cette longueur est aussi connue sous le nom de profondeur, $Depth$, du terme t. Cependant, à cause des multiples chemins dans un rDAG, il peut y avoir plusieurs profondeurs pour un seul terme. Dans la suite de la démonstration, nous définissons $Depth(t)$ comme une fonction qui associe à un terme t sa profondeur maximale afin de conserver la valeur correspondant à la spécificité maximale du terme :

$$Depth(t) = Max_i(|\Phi_i|) - 1 \ \mid \ \Phi_i \in RootPath(t). \tag{7}$$

Puisque $RootPath(Root)$=$\{\{\text{Root}\}\}$, on vérifie que $Depth(Root) = |\{\text{Root}\}| - 1 = 0$.
Après cela, nous définissons la fonction $Ancestors$ afin de récupérer les termes ancêtres d'un terme t donné, comme les éléments α d'un chemin $\Phi \in RootPath(t)$.

$$Ancestors : T \longrightarrow \mathcal{P}(T)$$

et

$$Ancestors(t) = \{\alpha \in T | \exists \Phi, (\Phi \in RootPath(t)) \wedge (\alpha \in \Phi)\}. \tag{8}$$

Par conséquent, les ancêtres communs de deux termes t_a and t_b peuvent se définir comme :

$$CommonAnc(t_a, t_b) = Ancestors(t_a) \cap Ancestors(t_b). \tag{9}$$

Soit $LCAset(t_a, t_b)$ l'ensemble des ancêtres communs les plus spécifiques des termes t_a, t_b. Ce sont les ancêtres communs qui sont à une distance maximale du nœud racine du graphe de l'ontologie. En d'autres termes, leur profondeur est la plus grande profondeur des termes $\alpha \in CommonAnc(t_a, t_b)$. Cette distance est une valeur unique mais peut correspondre à plus d'un terme LCA :

$$LCAset : T\text{x}T \longrightarrow \mathcal{P}(T) \ ,$$

$$\begin{aligned} LCAset(t_a, t_b) = \{\alpha \in CommonAnc(t_a, t_b) \mid \\ Depth(\alpha) = Max_i(Depth(a_i)), \\ a_i \in CommonAnc(t_a, t_b)\}. \end{aligned} \tag{10}$$

Ayant défini $LCAset$, il est possible de définir un sous-ensemble de chemins à partir du terme $Root$ vers un terme t et qui passent par un LCA, et par conséquent incluent les chemins les plus longs entre ce terme $Root$ et le LCA. Nous nommons ces chemins $ConstrainedRootPath$, et de fait, ils peuvent être calculés pour toute paire (t, s) avec $s \in Ancestors(t)$:

$$\begin{aligned} ConstrainedRootPath(t, s) = \{\Phi_i \in RootPath(t) \mid \\ (s \in \Phi_i) \wedge \\ (\forall \Phi_j, ((\Phi_j \in RootPath(s)) \wedge \\ (\Phi_j \subset \Phi_i)) \Rightarrow (|\Phi_j| = Depth(s) + 1))\}. \end{aligned} \tag{11}$$

Ceci nous ramène à poser la définition de la longueur du chemin ("Path Length") notée $PL_k(t, s)$, pour $s \in Ancestors(t)$ et pour un chemin $\Phi_k \in ConstrainedRootPath(t, s)$ comme :

$$PL_k(t, s) = |\Phi_k| - 1 - Depth(s). \tag{12}$$

Pour un $LCA \in LCAset(t_i, t_j)$, nous pouvons définir le plus court chemin (shortest path length : SPL) entre deux termes t_i et t_j qui passent par un LCA par la formule suivante :

$$\begin{aligned} SPL(t_i, t_j, LCA) = Min_k(PL_k(t_i, LCA)) + \\ Min_h(PL_h(t_j, LCA)). \end{aligned} \tag{13}$$

Le plus petit SPL entre les termes t_i et t_j est alors calculé en considérant tous leurs LCAs possibles par :

$$\begin{aligned} MinSPL(t_i, t_j) = Min_l(SPL(t_i, t_j, LCA_l)) \mid \\ LCA_l \in LCAset(t_i, t_j). \end{aligned} \tag{14}$$

Maintenant, si nous reprenons la définition de la formule du produit scalaire entre vecteurs de la base (formule (4)), nous pouvons relier la somme $Depth(t_i)+Depth(t_j)$ dans le dénominateur de cette équation aux expressions $MinSPL(t_i, t_j)$ et $Depth(LCA)$. A partir de la formule (7) nous avons $Depth(t_i) = Max_k(|\Phi_k|-1)$, avec $\Phi_k \in RootPath(t_i)$. Or de la formule (12) nous déduisons que avons $|\Phi_k| - 1 = Depth(LCA) + PL_k(t_i, LCA)$ avec $\Phi_k \in ConstrainedRootPath(t_i, LCA)$ et $ConstrainedRootPath(t_i, LCA) \subset RootPath(t_i)$. Pour des termes t_i, t_j, et un $LCA \in LCAset(t_i, t_j)$, il sera donc facile de démontrer que :

$$Depth(t_i) \geq Min_k(PL_k(t_i, LCA)) + Depth(LCA). \tag{15}$$

De façon similaire,

$$Depth(t_j) \geq Min_h(PL_h(t_j, LCA)) + Depth(LCA). \tag{16}$$

Par conséquent,

$$\begin{aligned} Depth(t_i) + Depth(t_j) \geq \\ SPL(t_i, t_j, LCA) + 2 * Depth(LCA) \geq \\ MinSPL(t_i, t_j) + 2 * Depth(LCA). \end{aligned} \tag{17}$$

Rappelons que dans le cas d'un arbre, cette double égalité est une simple égalité car il n'y a qu'un seul $SPL(t_i, t_j)$.
La similarité sémantique entre deux termes est inversement proportionnelle à la longueur du chemin séparant les deux termes de leurs LCA. Quand nous adaptons la formule (4) en remplaçant dans le dénominateur la somme $Depth(t_i) + Depth(t_j)$ par une somme plus petite $MinSPL(t_i, t_j) + 2 * Depth(LCA)$, nous assurons que le produit scalaire entre les vecteurs de la base soit maximisé. Avec cette adaptation, le produit scalaire *IntelliGO* correspondant aux deux termes GO t_i et t_j est défini par :

$$\vec{e_i} * \vec{e_j} = \frac{2 * Depth(LCA)}{MinSPL(t_i, t_j) + 2 * Depth(LCA)}. \tag{18}$$

Il est facilement vérifiable qu'avec cette définition, le produit scalaire est dans l'intervalle [0,1]. Nous pouvons observer d'ailleurs pour $i = j$, $\vec{e_i}$*$\vec{e_i} = 1$, puisque $MinSPL(t_i, t_j) = 0$. De plus, quand deux termes sont reliés uniquement par le nœud racine du rDAG, nous avons $\vec{e_i}$*$\vec{e_i} = 0$ car $Depth(Root) = 0$. Dans tous les autres cas de figures, la valeur prise par le produit scalaire est une valeur non nulle d'une similarité sémantique entre deux termes (*edge-based*). Il faut noter par ailleurs que cette valeur dépend clairement de la structure du rDAG de l'ontologie GO.

1.2.3 Définition de la mesure de similarité fonctionnelle IntelliGO entre deux produits de gènes

La mesure de similarité sémantique *IntelliGO* entre deux gènes g et h représentés par leurs vecteurs $\vec{g}$ et $\vec{h}$, respectivement est définie comme l'adaptation de la formule du cosinus généralisé de Ganesan *et al.* [GGMW03], selon :

$$SIM_{IntelliGO}(g, h) = \frac{\vec{g} * \vec{h}}{\sqrt{\vec{g} * \vec{g}}\sqrt{\vec{h} * \vec{h}}}, \tag{19}$$

Où :

- $\vec{g} = \sum_i \alpha_i * \vec{e_i}$: la représentation vectorielle du gène g dans l'espace vectoriel de la mesure *IntelliGO*.
- $\vec{h} = \sum_j \beta_j * \vec{e_j}$: la représentation vectorielle du gène h dans l'espace vectoriel de la mesure *IntelliGO*.
- $\alpha_i = w(g, t_i)$*$IAF(t_i)$: le coefficient du terme t_i pour le gène g, où $w(g, t_i)$ représente le poids assigné au code d'évidence entre le terme t_i et le gène g. Le coefficient $IAF(t_i)$ est la fréquence inverse (inverse annotation frequency) du terme t_i.
- $\beta_j = w(h, t_j)$*$IAF(t_j)$: le coefficient du terme t_j pour le gène h.
- $\vec{g}$*$\vec{h}$=$\sum_{i,j} \alpha_i * \beta_j$*$\vec{e_i} * \vec{e_j}$: représente le produit scalaire entre les deux vecteurs des gènes g et h.
- $\vec{e_i} * \vec{e_j} = \frac{2 * Depth(LCA)}{MinSPL(t_i, t_j) + 2 * Depth(LCA)}$: représente le produit scalaire entre $\vec{e_i}$ et $\vec{e_j}$ ($\vec{e_i}$*$\vec{e_j} \neq 0$ si les termes t_i et t_j partagent des ancêtres communs autres que la racine du rDAG).

Notons que dans l'espace vectoriel associé à la mesure *IntelliGO*, les produits scalaires $\vec{g} * \vec{g}$ et $\vec{h} * \vec{h}$ ne sont pas réduits à la somme des carrés des coefficients des termes d'annotation des gènes g et h du fait de la non nullité des produits scalaires $\vec{e_i} * \vec{e_j}$.

1.3 Mise en œuvre et algorithme

L'algorithme de la mesure *IntelliGO* a été conçu pour calculer la similarité de deux produits de gènes, en prenant comme entrée leurs identifiants $NCBI$ de la base GENE [11], et comme paramètres un aspect de l'ontologie GO c'est à dire BP ou MF ou CC, une espèce du monde vivant, et une liste de poids associés aux codes d'évidence. Le résultat de l'algorithme est une valeur de similarité entre les deux gènes en entrée.
Pour mettre en œuvre la mesure de similarité, une première étape de préparation des données d'annotations est nécessaire. J'ai utilisé un fichier d'annotation du NCBI [NCB], qui contient une liste non redondante de termes d'annotation GO auxquels sont associés les gènes qu'ils annotent ainsi que le code d'évidence qui décrit la façon dont l'annotation a été faite. Les valeurs des coefficients IAF sont calculées pour chaque terme selon son occurrence dans ce corpus en utilisant la formule (2), selon une espèce donnée. Les résultats sont alors stockés dans une table (Terme GO $\times$ IAF) qui est sauvegardée dans un fichier *SpeciesIAF_File*. Ensuite, j'ai utilisé le schéma relationnel sous lequel est disponible l'ontologie GO. J'ai construit à partir du fichier d'annotation NCBI, toutes les paires de termes GO possibles, pour interroger l'outil AMIGO [AMI] afin de récupérer les valeurs des LCA, $Depth(LCA)$ et SPL. Ces traitements se font par des requêtes SQL spécifiques à la base de données de GO. Ayant ces valeurs, le produit scalaire entre chaque couple de termes GO sera donc calculé en utilisant la formule (18). Les résultats sont alors sauvegardés dans un fichier spécifique *DotProduct_File*.
L'algorithme de la mesure de similarité *IntelliGO* commence par le filtrage du fichier d'annotation NCBI selon les trois paramètres choisis par l'utilisateur, c'est à dire un aspect de l'ontologie (BP, CC ou MF), une espèce (Humain, Levure, Souris, ...), et une liste de poids qui seront associés aux différents codes d'évidence. Après cette étape, les annotations sont filtrées pour ne garder que les gènes et leurs termes GO choisis selon ces trois paramètres, et qui seront sauvegardés dans un fichier *CuratedAnnotation_File*. Si un gène est annoté par le même terme GO mais qualifié avec plusieurs codes d'évidence, l'algorithme garde seulement le code avec le poids le plus grand qui est donné en paramètre dans la liste des poids. Après ceci, ayant deux identifiants NCBI de gènes, la similarité fonctionnelle *IntelliGO* peut être calculée selon les étapes de l'algorithme (1).

2 Evaluation de la mesure de similarité sémantique IntelliGO

2.1 Présentation des collections de gènes pour l'évaluation

Nous avons évalué notre mesure *IntelliGO* avec deux protocoles différents et en utilisant des jeux de données pris dans deux espèces très différentes et impliquant deux aspects distincts de l'ontologie GO. Au total quatre collections d'ensembles de gènes sont utilisées (voir tables 10 et 11).
D'une part, j'ai choisi 13 voies métaboliques (pathways) de l'espèce levure et 13 autres pathways de l'espèce humaine. Ces pathways, disponibles dans la base KEGG [KEG, KGF$^+$10], contiennent un nombre raisonnable de gènes (entre 10 et 30). Les pathways sélectionnés sont présentés dans la Table (10). Les gènes qui sont dans les pathways ont été extraits en utilisant l'outil d'extraction *DBGET* de la base KEGG [DBG]. Les gènes appartenant à un même pathway sont supposés être reliés à des processus biologiques similaires, par conséquent les similarités fonctionnelles *IntelliGO* calculées sur ces deux jeux de données devraient l'être en choisissant l'aspect BP de GO.

11. http ://www.ncbi.nlm.nih.gov/

Algorithm 1 Algorithme de calcul de la similarité fonctionnelle *IntelliGO* entre deux gènes.

Require: Deux identifiants NCBI des gènes g_1 et g_2, une Liste de poids des codes d'évidence *List*, un aspect de l'ontologie *GO_Aspect*, une espèce donnée *Species*, un fichier d'annotation *Annotation_File*, un fichier des contenus en information des annotations selon l'espèce q *SpeciesIAF_File*, un fichier *DotProduct_File* contenant le produit scalaire entre tous les termes GO présents dans *Annotation_File*.

Ensure: $SIM_{IntelliGO}(g_1,g_2)$: Une valeur de similarité fonctionnelle.

1: Créer à partir du fichier *Annotation_File*, le fichier *CuratedAnnotation_File* ne contenant que les annotations des gènes de l'espèce et selon l'aspect GO choisi.
2: Extraire à partir du fichier *CuratedAnnotation_File* la liste des termes GO annotant les deux gènes et leurs codes d'évidence associés.
3: Calculer à partir du fichier *SpeciesIAF_File* et à partir de la liste des poids *List* les coefficients de pondération des deux gènes dans l'espace vectoriel d'*IntelliGO*.
4: Récupérer à partir du fichier *DotProduct_File* la similarité sémantique entre toutes les paires de termes GO qui annotent les deux gènes.
5: Calculer la valeur de similarité *IntelliGO* en utilisant la formule (19)
6: **return** $SIM_{IntelliGO}(g_1,g_2)$.

D'autre part, j'ai utilisé les clans de protéines de la base Pfam [pfa, FMSB+] pour construire deux jeux de données de séquences humaines et levure. Les clans de ces deux jeux de données sont disponibles dans la Table 11. Pour chaque clan Pfam, nous avons utilisé toutes les entrées Pfam afin d'interroger la base Uniprot [uni] pour extraire les identifiants de gènes humains et de la levure. L'hypothèse posée sur ces deux collections d'ensembles de gènes suppose que les gènes partageant des domaines similaires dans un clan Pfam, partagent aussi des fonctions moléculaires similaires, et par conséquent les similarités fonctionnelles devraient être calculées selon l'aspect MF de l'ontologie GO.

Nous avons étudié la distribution des codes d'évidence dans le corpus NCBI utilisé selon les annotations de type BP et MF, et en considérant les deux espèces humaine et levure. Les histogrammes sont affichés dans la Figure 15. Les nombres d'annotations assignées à un gène selon un code d'évidence sont représentés dans une barre d'histogramme pour chaque code d'évidence. Le nombre d'annotations non-IEA concernant l'espèce levure est 18 496 pour l'aspect BP, et 9564 pour l'aspect MF. 21 462 et 16 243 annotations non-IEA sont dénombrées pour l'espèce humaine en considérant les aspects BP et MF, respectivement.

2.2 Protocole d'évaluation

Nous considérons S une collection de p ensembles de gènes où $S=\{S_1, S_2, ..., S_p\}$, où un ensemble S_k peut être un pathway KEGG ou un clan Pfam. Chaque ensemble $S_k=\{g_{k1}, g_{k2}, ..., g_{kn}\}$ regroupe l'ensemble de n gènes qui sont présents dans S_k. Nous posons $Sim(g, h)$ une similarité fonctionnelle entre deux gènes g et h. Nous avons proposé un protocole d'évaluation de la mesure de similarité fonctionnelle *IntelliGO*, qui est appliqué sur les 4 collections d'ensembles de gènes présentées précédemment. Ce protocole est applicable à toute mesure de similarité fonctionnelle Sim.

La première étape du protocole consiste à calculer pour chaque ensemble de gènes dans une des quatre collections, une similarité moyenne entre tous les gènes se trouvant dans l'ensemble considéré S_k. Cette similarité *Intra-set* se définit pour un ensemble de gènes S_k d'une collection

Catégorie KEGG	Sous catégorie KEGG	Pathway Levure	Nom	Nb gènes	Pathway Humain	Nom	Nb
01100 Metabolism	01101 Carbohydrate Metabolism	sce00562	Inositol phosphate metabolism	15	hsa00040	Pentose and glucuronate interconversions	
	01102 Energy Metabolism	sce00920	Sulfur metabolism	13	hsa00920	Sulfur metabolism	
	01103 Lipid Metabolism	sce00600	Sphingolipid metabolism	13	hsa00140	C21-Steroid homone metabolism	
	01105 Amino Acid Metabolism	sce00300	Lysine biosynthesis	13	hsa00290	Valine, leucine and isoleucine biosynthesis	
		sce00410	Alanine biosynthesis	8			
	01107 Glycan Biosynthesis and Metabolism	sce00514	O-Mannosyl glycan biosynthesis	13	hsa00563	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	
	01109 Metabolism of Cofactors and Vitamins	sce00670	One carbon pool by folate	14	hsa00670	One carbon pool by folate	
	01110 Biosynthesis of Secondary Metabolites	sce00903	Limonene and pinene degradation	7	hsa00232	Caffeine metabolism	
01120 Genetic Information Processing	01121 Transcription	sce03022	Basal transcription factors	24	hsa03022	Basal transcription factors	
					hsa03020	RNA polymerase	
	01123 Folding, Sorting and Degradation	sce04130	SNARE interactionst in vesicular transpor	23	hsa04130	SNARE interactions in vesicular transport	
	01124 Replication and Repair	sce03450	Non-homologous end-joining	10	hsa03450	Non-homologous end-joining	
					hsa03430	Mismatch repair	
01130 Environmental Information Processing	01132 Signal Transduction	sce04070	Phosphatidylinositol signaling system	15			
01140 Cellular Processes	01151 Transport and Catabolism	sce04140	Regulation of autophagy	17			
01160 Human Diseases	01164 Metabolic Disorders				hsa04950	Maturity onset diabetes of the young	
Nombre total de gènes				185			
Rapport Non-IEA :IEA				572 :435 (1.3)			56

TABLE 10 – Description des 26 jeux de données relatifs aux pathways KEGG. Les catégories et les sous-catégories KEGG sont indiquées pour chaque pathway, ainsi que son nom et le nombre de gènes qu'il contient (KEGG version Dec 2009). Le rapport non-IEA :IEA réfère aux annotations GO de type *Biological Process* pour chaque ensemble de gènes dans chaque espèce.

S par la formule suivante :

$$Intra\_Set\_Sim(S_k) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} Sim(g_{ki}, g_{kj})}{n^2}. \quad (20)$$

Chaque ensemble S_k d'une collection S regroupe des gènes qui, selon l'expert du domaine, partagent des fonctions biologiques similaires. Dans le cas idéal, une valeur *Intra-set* élevé pour une similarité sémantique donnée Sim sur un ensemble S_k devrait refléter la cohésion élevée des gènes de cet ensemble.
Par ailleurs, étant donné deux ensembles de gènes S_k et S_l composés de n et m gènes respectivement, une similarité moyenne *Inter-set* entre les gènes de ces deux ensembles est calculée, selon la formule suivante :

$$Inter\_Set\_Sim(S_k, S_l) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} Sim(g_{ki}, g_{lj})}{n * m}. \quad (21)$$

Notons que Sim est une fonction qui prend des valeurs de similarité fonctionnelle dans l'intervalle [0,1], dans lequel sont définies aussi les valeurs prises par les fonctions $Intra\_Set\_Sim$ et $Inter\_Set\_Sim$.
Enfin, un pouvoir de discrimination (DP) est calculé pour chaque ensemble de gène S_k d'une collection S, comme étant un ratio entre les valeurs intra-set et inter-set. La valeur DP est comprise dans l'intervalle $[0,+\infty]$, et reflète la capacité d'une mesure à différencier entre deux ensembles différents de gènes. Plus les valeurs DP sont élevées, plus la mesure de similarité sémantique

Accession Pfams clan (Levure)	Nb gènes	Nom Pfams clan	Accession Pfams clan (humain)	Nb gènes	Nom Pfams clan
CL0328.1	15	2heme_cytochrom	CL0099.10	18	ALDH-like
CL0059.12	13	6_Hairpin	CL0106.10	8	6PGD_C
CL0092.9	8	ADF	CL0417.1	9	BIR-like
CL0099.10	11	ALDH-like	CL0165.8	5	Cache
CL0179.11	11	ATP-grasp	CL0149.9	7	CoA-acyltrans
CL0255.6	7	ATP_synthase	CL0085.11	12	FAD_DHS
CL0378.1	10	Ac-CoA-synth	CL0076.9	18	FAD_Lum_binding
CL0257.6	18	Acetyltrans-like	CL0289.3	6	FBD
CL0034.12	11	Amidohydrolase	CL0119.10	7	Flavokinase
CL0135.8	14	Arrestin_N-like	CL0042.9	10	Flavoprotein
Nombre total de gènes	118			100	
Rapport Non-IEA :IEA	121 :366 (0.3)			144 :309 (0.46)	

TABLE 11 – Description des 20 jeux de données relatifs aux clans Pfam. Les Clans sont indiqués par leur identifiant d'accession dans la base Sanger Pfam (version Oct. 2009) et par leurs nombre de gènes extraits pour les deux espèces humaine (droite) et levure (gauche). Chaque clan contient plusieurs entrées du fichier Pfam_C disponible dans [pfa]. Le rapport non-IEA :IEA réfère aux annotations GO de l'aspect *Molecular Function* pour chaque ensemble de gène dans chaque espèce.

Sim est discriminante.

$$DP_{Sim}(S_k) = \frac{(p-1)Intra\_Set\_Sim(S_k)}{\sum_{i=1,i\neq k}^{p} Inter\_Set\_Sim(S_k, S_i)}. \quad (22)$$

Nous avons comparé les valeurs obtenues avec la mesure de similarité *IntelliGO* contre quatre mesures présentées dans l'état de l'art, à savoir : la mesure de *Lord* (équation 16), la mesure d'*Al-Mubaid* (équation 17), la mesure *SimGIC* (formule 20), et la mesure basée sur le *cosinus pondéré* (formule 21). Nous avons normalisé les valeurs obtenues avec la mesure de Lord puisque d'après l'équation (16), les résultats pouvaient être supérieurs à 1. Les mesures choisies dans cette évaluation représentent la diversité des techniques utilisées et exploitent plusieurs types d'information comme le plus court chemin entre annotations ou le contenu en information.
Pour chaque ensemble de gènes présent dans une collection, nous avons tout d'abord évalué la mesure de similarité *IntelliGO* en comparant les valeurs *Intra-set* obtenues avec cette mesure contre celles obtenues avec les quatre autres mesures de similarité considérées dans l'évaluation. Ensuite, nous avons étudié l'impact du choix de différentes valeurs affectées aux codes d'évidence pour notre mesure *IntelliGO*. Pour cela nous avons calculé les valeurs *Intra-set* en faisant varier les poids des codes d'évidence. La table 12 représente les quatre listes qui ont été utilisées à cet effet. La liste de poids qui sera gardée devrait être celle qui fournit les meilleures valeurs *Intra-set*. Par

	Auth		Exp						Comp						Cur		Auto
EC	TAS	NAS	EXP	IDA	IPI	IMP	IGI	IEP	ISS	RCA	ISA	ISO	ISM	IGC	IC	ND	IEA
Liste1	1																
Liste2	1	0.5	0.8	0.8	0.8	0.8	0.8	0.8	0.6	0.6	0.6	0.6	0.6	0.6	0.5	0	0.4
Liste3	1	0.5	0.8	0.8	0.8	0.8	0.8	0.8	0.6	0.6	0.6	0.6	0.6	0.6	0.5	0	0
Liste4	0																1

TABLE 12 – Les poids affectés aux EC sont listés dans les listes nommées de 1 à 4.

la suite du protocole d'évaluation, nous avons comparé les valeurs du pouvoir de discrimination (DP) obtenues avec *IntelliGO* et les quatre mesures de similarité considérées. Finalement, nous avons utilisé un outil en ligne qui s'appelle "Collaborative Evaluation of GO-based Semantic Similarity Measures (CESSM) tool" [ces], dont le principe sera détaillé par la suite. Les matrices de similarité sont calculées pour les mesures *IntelliGO*, Lord, cosinus pondéré et Al-Mubaid avec des programmes implémentés en C++, tandis que la mesure SimGIC est disponible dans le

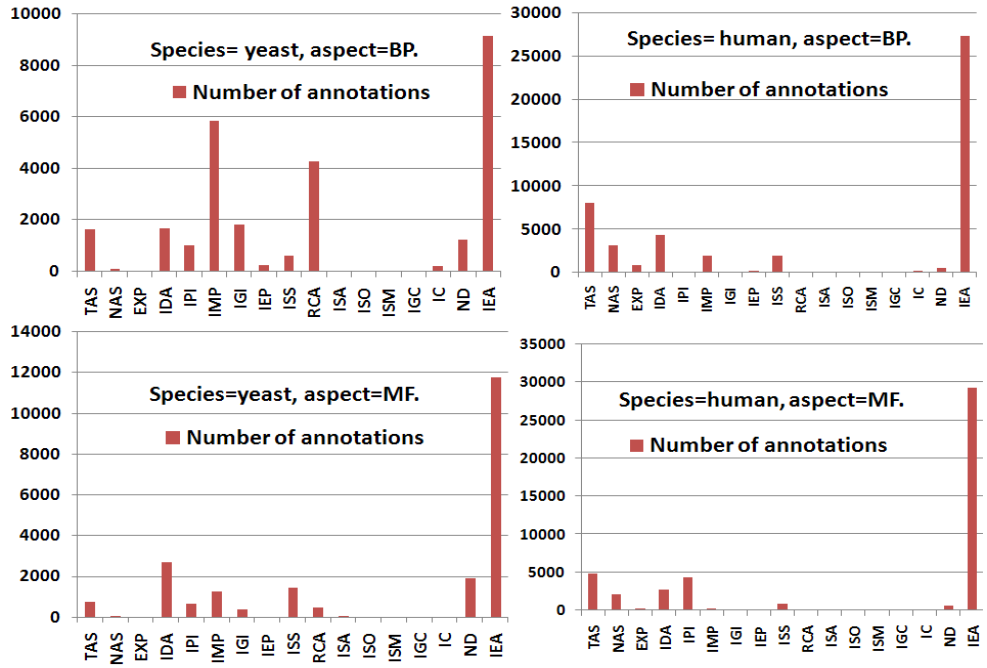


FIGURE 15 – Distribution des codes d'évidence pour les annotations dans la levure et l'espèce humaine en considérant les aspects BP et MF du fichier d'annotation NCBI utilisé dans cette étude (Nov. 2009).

package csbl.go [Ova]. Les méthodes d'évaluation (Intra-set, Inter-set, Discrimminative Power) ont été implémentées en C++. Une version en ligne de la mesure *IntelliGO* est disponible sur la plateforme du projet MBI du LORIA à Nancy : http ://plateforme-mbi.loria.fr/intelligo/

2.3 Mesure Intra-set

Nous avons produit toutes les valeurs *Intra-set* pour les mesures *IntelliGO*, Lord, Al-Mubaid, SimGIC et le cosinus pondéré. Concernant la mesure *IntelliGO* nous avons considéré la *Liste1* (tous les poids sont à 1), voir Table 12.
Les résultats qui concernent les pathways KEGG pour les deux espèces humaine et levure ("human" et "yeast" sur les figures) en utilisant les annotations BP sont dans la Figure 16. Pour chaque identifiant de pathway KEGG (axe des X), une valeur de similarité *Intra-set* est représentée comme un histogramme pour toutes les mesures de similarité comparées (axe des Y). Les valeurs *Intra-set* varient d'un ensemble à un autre, reflétant une variation de la cohérence des annotations des gènes dans les pathways. D'après la Figure 16, nous pouvons observer des variations uniformes d'un pathway à un autre pour toutes les mesures à l'exception de celle de Lord. Par exemple, les valeurs *Intra-set* pour le pathway sce00410 sont plus petites que le pathway sce00300 pour toutes les mesures à l'exception de la mesure de Lord. Le même cas est observé entre les pathways hsa00920 et hsa00140. Un avantage de la mesure *IntelliGO* est observé par rapport aux autres mesures puisque toutes les valeurs *Intra-set* sont supérieures ou égales à 0,5. Les valeurs relativement faibles observées pour la mesure du cosinus pondéré peuvent être expliquées par des paires de produits scalaires de valeur nulle obtenues avec cette mesure de similarité. Ce phénomène est causé par le fait que cette mesure suppose que les dimensions de l'espace vectoriel sont orthogonales l'une à l'autre. De ce fait, l'absence d'annotations communes

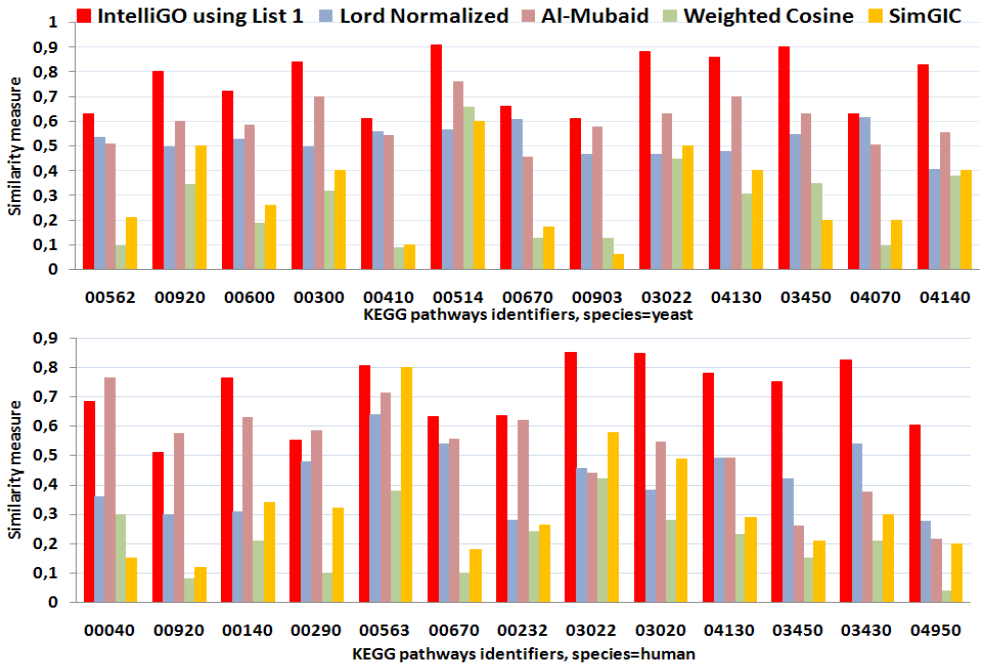


FIGURE 16 – Valeurs de similarité Intra-set avec les pathways KEGG en ne considérant que les annotations BP. La similarité Intra-set est calculée comme la moyenne des similarités entre paires de gènes présents dans un pathway KEGG, avec les mesures *IntelliGO* (en utilisant *Liste1* des poids des codes d'évidences), Lord-normalisée, Al-Mubaid, SimGIC et le Cosinus pondéré. Une collection de 13 pathways a été sélectionnée de la base KEGG pour l'espèce levure (haut de la figure), et l'espèce humaine (en bas de la figure).

à deux gènes produit un produit scalaire nul, et par conséquent une valeur de similarité nulle entre les deux gènes. De fait, des paires de valeurs de similarité nulles sont observées dans tous les pathways sauf un seul dans l'espèce humaine et trois pour la levure.
Des résultats très similaires sont obtenus en utilisant les clans Pfams qui sont présentés dans les collections 3 et 4 (voir Figure 17). Dans ces deux collections d'ensembles, avec la mesure *IntelliGO* nous obtenons aussi des valeurs *Intra-set* supérieures ou égales à 0,5, ce qui n'est pas forcément le cas pour les autres mesures de similarité. Les mêmes remarques que dans le cas des pathways se font aussi pour certaines valeurs très faibles obtenues avec la même mesure du cosinus pondéré, pour les mêmes raisons que celles évoquées ci-dessus. Pour ces raisons, nous avons décidé de poursuivre les évaluations sans tenir compte de cette dernière mesure de similarité.
En résumé, notre comparaison en utilisant les valeurs *Intra-set* comme indice d'évaluation sur les quatre collections d'ensembles de gènes, illustre parfaitement la robustesse de la mesure de similarité *IntelliGO* quant à son aptitude à capturer la cohésion interne des annotations fonctionnelles au sein d'ensembles de gènes, prédéfinis en tenant compte des relations sémantiques entre ces annotations.

2.4 Influence des poids des codes d'évidences sur la mesure IntelliGO

La seconde partie de notre évaluation, concerne l'étude de l'impact des poids affectés aux codes d'évidence sur la mesure de similarité *IntelliGO*. Comme première expérience, nous avons utilisé les quatre listes de poids présentées dans la Table 12. Dans la *Liste1*, tous les poids sont

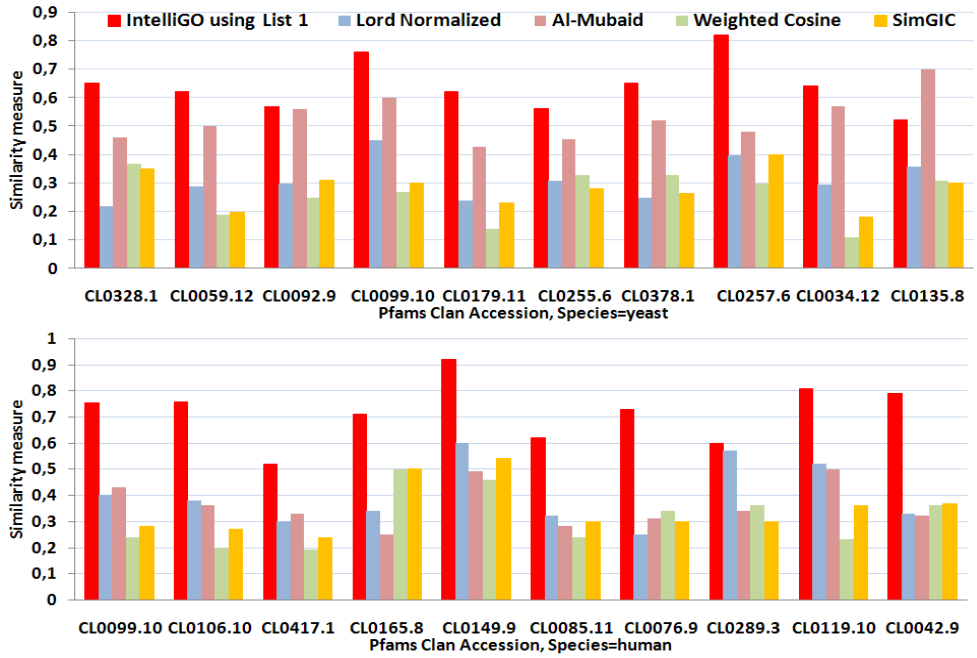


FIGURE 17 – Valeurs de similarité Intra-set en considérant les clan Pfam et en ne gardant que les annotations MF. La similarité Intra-set est calculée comme la moyenne des similarités entre paires de gènes présents dans un clan Pfam, avec les mesures *IntelliGO* (en utilisant *Liste1* des poids des codes d'évidences), Lord-normalisée, Al-Mubaid, SimGIC et le Cosinus pondéré. Une collection de 10 clans de protéines ont été sélectionnés de la base Pfam Sanger pour l'espèce levure (haut de la figure), et l'espèce humain (en bas de la figure).

à 1, ce qui veut dire que la contribution des codes d'évidence aux valeurs de similarité *IntelliGO* est la même. C'est pour cela que nous avons utilisé cette liste dans la phase d'évaluation initiale avec les autres mesures de similarité sur l'indice *Intra-set* (Figures 16 et 17), puisque les autres mesures ne considèrent pas de façon variée les codes d'évidence des annotations. Dans la *Liste2*, les poids ont été choisis arbitrairement pour représenter l'hypothèse que la catégorie des codes *Exp* est plus fiable que ceux de la catégorie *Comp*, et que les codes non supervisés *IEA* sont moins fiables que les codes de la catégorie *Comp*. La *Liste3* exclut carrément le code *IEA* afin de tester la mesure de similarité *IntelliGO* en ne gardant que les annotations plus ou moins validées. Finalement, la *Liste4* représente le cas opposé de la *Liste3* en ne gardant que les annotations *IEA* pour évaluer aussi leur impact sur la mesure de similarité *IntelliGO*.
Ces quatre listes de poids sont utilisées pour calculer les valeurs *Intra-set* avec la mesure de similarité *IntelliGO* en utilisant les quatre collections d'ensembles de gènes. Pour chaque collection, les distributions des valeurs (intervalles de taille 0,2) des similarité entre paires de gènes sont représentées sous forme d'histogramme (18).
A la gauche de chaque histogramme, une barre *Missing Values* (MV) représente le nombre de similarités entre paires de gènes dont les valeurs ne peuvent pas être calculées avec la *Liste3* ou *Liste4* à cause de l'absence totale d'annotations pour certains gènes. Comme démontré auparavant, les valeurs *Intra-set* de la mesure *IntelliGO* sont supérieures à 0,5 (avec *Liste1*), par conséquent les nombres les plus élevés qui concernent les similarités entre paires de gènes sont observés dans les intervalles 0,6-0,8 et 0,8-1,0 pour toutes les listes de poids utilisées. Pour les

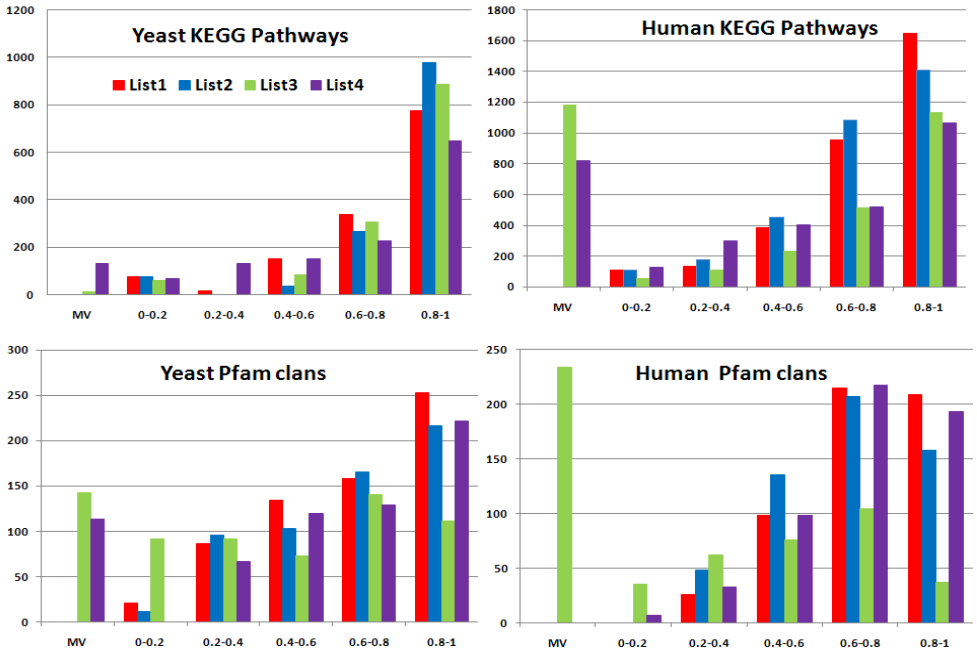


FIGURE 18 – Etude de l'influence des différents poids affectés aux codes d'évidence. Quatre listes de poids ont été utilisées, en calculant la distribution des similarités entre paires de gènes dans le calcul des valeurs *Intra-set*. Les pathways KEGG sont utilisés en considérant les annotations BP, tandis que les gènes des clans Pfam sont considérés avec les annotations MF. La barre MV (Valeurs Manquantes : *Missing Values*) dans les histogrammes représente le nombre de valeurs de similarité entre paires de gènes dont le calcul est impossible quand la *Liste3* ou la *Liste4* sont utilisées, du fait de l'absence d'annotations de certains gènes. Les intervalles des valeurs de similarité entre paires de gènes sont affichés dans l'axe des x des histogrammes, tandis que l'axe y représente le nombre de valeurs de similarité entre paires de gènes appartenant à chaque intervalle.

Liste1 et *Liste2*, la distribution des valeurs de similarité semble similaire pour tous les ensembles de gènes des quatre collections. L'exclusion des annotations de type IEA dans la *Liste3* ou le fait de les prendre exclusivement dans la *Liste4* ont des effets différents selon que l'on teste les pathways KEGG ou les clans Pfam, c'est-à-dire selon que l'on considère les annotations BP ou MF. Les effets du choix des *Liste3* ou *Liste4* sont aussi différents selon que l'on teste les jeux de données humain ou levure, ce qui reflète la différence des ratios entre les annotations IEA contre les non-IEA dans ces deux espèces (Figure 15, et Tables 10 et 11). Concernant les pathways KEGG, la variation la plus remarquable est observée avec la *Liste4* qui génère un nombre de valeurs de similarité plus faible que les autre listes dans la classe d'intervalle 0,8-1, et un nombre significatif de valeurs manquantes (barre MV) dans la levure. Ceci pourrait être expliqué par le fait que sur ce jeu de données, l'utilisation exclusive des annotations BP de type IEA génère des valeurs de similarité significativement faibles et exclut certains gènes qui n'ont pas d'annotations IEA (au total 11 gènes chez la levure). Ceci reflète aussi le ratio relativement élevé (1,3) des annotations non-IEA par rapport aux IEA pour les annotations BP dans ce jeu de données. Un comportement très similaire est observé dans les pathways KEGG de l'espèce humaine, non seulement en considérant *Liste4* mais aussi *Liste3*. Un nombre élevé de valeurs

manquantes (barre MV) dans cette collection d'ensembles de gènes est causé par le nombre élevé de gènes n'ayant pas d'annotations BP de type IEA (49 gènes) ou ayant seulement des annotations BP de type IEA (68 gènes). Ce type d'analyse démontre que sur cette collection d'ensembles de gènes, les annotations IEA sont importantes pour capturer les valeurs *Intra-set* pendant le processus de calcul des similarités sémantiques entre paires de gènes, mais qu'elles ne sont pas suffisantes quand elles sont utilisées toutes seules. Pour les collections issues des clans Pfam chez la levure et dans l'espèce humaine, la distribution des valeurs obtenues avec la *Liste3* est clairement dégradée dans les classes d'intervalle de 0,6 à 1 et elle génère beaucoup de valeurs manquantes (barre MV). Ceci reflète aussi que les annotations MF de type IEA sont importante pour capturer les valeurs *Intra-set* dans ces collections de données. En effet, les ratios entre les annotations non-IEA et IEA sont de l'ordre de 0,3 et 0,46 pour la levure et l'espèce humaine pour les clans Pfam respectivement (Table 11). Au total, 19 gènes sont annotés seulement avec des codes IEA dans la levure et 29 chez l'humain. Concernant *Liste4*, en utilisant seulement les annotations MF de type IEA, elle n'apporte pas un grand changement par rapport à *Liste1* et *Liste2*. Ceci suggère que ces annotations sont suffisantes pour capturer les valeurs *Intra-set* lors du processus de calcul des paires de similarité sémantique. Cependant, un nombre significatif de valeurs manquantes sont observées dans les clans Pfam chez la levure, avec 20 gènes manquant d'annotations MF de type IEA.
En résumé, l'utilisation des listes de poids variables pour les codes d'évidence dans le calcul de similarité fonctionnelle *IntelliGO* est intéressante pour mettre en évidence la contribution et l'impact de certains types d'annotations. Il faudrait signaler cependant que cette contribution dépend clairement du jeu de données utilisé et de l'aspect de l'ontologie GO considéré. D'autres listes de poids pourraient être utilisées si l'utilisateur voulait mettre en avant l'effet de certains codes d'évidence particuliers en relation avec le jeu de données utilisé. Dans la suite de cette étude nous avons décidé de continuer notre analyse en utilisant *Liste2* qui nous semble la mieux adaptée puisqu'elle reflète la confiance variable accordée aux différents codes d'évidence pour les annotations des gènes.

2.5 Pouvoir de discrimination d'une mesure de similarité

La troisième partie de notre évaluation concerne le test du pouvoir de discrimination (DP) de la mesure de similarité *IntelliGO* comparée aux trois autres mesures gardées pour le reste de l'évaluation, à savoir : Lord, Al-Mubaid, SimGIC. Le calcul de la valeur DP (équation 22) illustre ici le pouvoir d'une mesure de similarité de distinguer entre deux ensembles de gènes fonctionnellement différents. Les résultats d'évaluation avec les valeurs DP sont présentés pour les quatre collections, dans les Figures 19 et 20. Pour les pathways KEGG en considérant les annotations BP (Figure 19), la mesure de similarité *IntelliGO* produit des valeurs DP supérieures ou égales à $1,3$ pour chaque pathway testé, et un maximum de $2,43$ est atteint pour le pathway *hsa*04130. D'autre part, les valeurs DP obtenues pour la mesure de Lord varient aux alentours de 1, plus spécialement avec les pathways de la levure, ce qui n'est pas une valeur discriminante. Les mesures d'Al-Mubaid et SimGIC génèrent quant à elles des valeurs hétérogènes qui varient entre $1,0$ et $2,5$, et $0,2$ et $2,3$, respectivement. Cette hétérogénéité indique que le pouvoir de discrimination de ces mesures n'est pas aussi stable que celui de la mesure *IntelliGO*. Des résultats très similaires sont obtenus avec la mesure *IntelliGO* en utilisant les clans Pfam avec les annotations MF, comme illustré dans la Figure 20. Dans ce cas, toutes les valeurs DP de la mesure *IntelliGO* sont supérieures à $1,5$, avec un maximum de $5,4$ pour le clan $CL0255.6$. Les trois autres mesures génèrent des valeurs DP peu élevées (par exemple la mesure de Lord pour les clans de la levure), ou très hétérogènes (le reste des valeurs). En résumé, ces résultats démontrent que la mesure

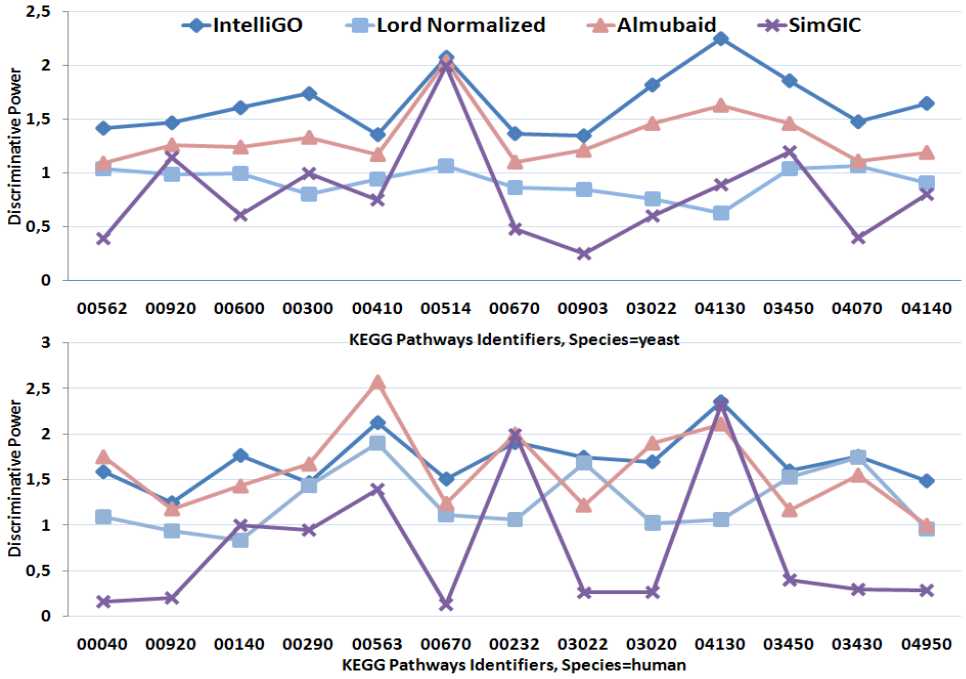


FIGURE 19 – Comparaison des valeurs du pouvoir de discrimination (DP) de quatre mesures de similarité en utilisant les pathways KEGG sur les annotations BP. Les valeurs DP obtenues avec *IntelliGO*, Lord-normalisée, Al-Mubaid, et SimGIC sont affichées pour chaque pathway KEGG pour la levure (Haut) et pour l'humain (Bas).

de similarité sémantique *IntelliGO* a une bonne capacité pour discriminer entre deux ensembles de gènes différents. Ceci peut être considéré comme une indication quant à l'utilisation de cette mesure dans un processus de classification fonctionnelle de gènes.

2.6 Evaluation avec l'outil CESSM

Une évaluation complémentaire a été exécutée en utilisant un outil très récent : *Collaborative Evaluation of GO-based Semantic Similarity Measures* (CESSM). Cet outil disponible en ligne depuis 2010 [ces] rend possible la comparaison d'une mesure de similarité sémantique donnée, avec des mesures déjà publiées, en utilisant la corrélation entre la similarité sémantique et la similarité de séquence de gènes., la similarité de composition en domaines Pfam, et la similarité de classification (EC) des enzymes [PPFC09]. L'outil utilise un jeu de données composé de 13430 paires de protéines contenant 1039 différentes protéines de diverses espèces. Ces paires de protéines sont caractérisées par leur similarité de séquence, le nombre de domaines Pfam en commun, et leur degré de relation dans la classification d'Enzyme, définissant ainsi trois indicateurs biologiques, appelés métriques *SeqSim*, *Pfam* et *ECC* respectivement. Les valeurs de similarité sémantique calculées selon plusieurs approches sont analysées en corrélation avec ces trois indicateurs biologique. L'utilisateur est tout d'abord invité à télécharger le jeu de données proposé dans l'outil CESSM, pour calculer les paires de similarité sémantique avec sa propre mesure de similarité. L'outil analyse les résultats et retourne un graphe et un tableau qui affiche les coefficients de corrélation de Pearson, calculés entre les valeurs de similarité de la mesure de l'utilisateur ainsi que 11 autres mesures de similarité disponibles.
Nous présentons les résultats concernant les valeurs de corrélation retournés par l'outil CESSM

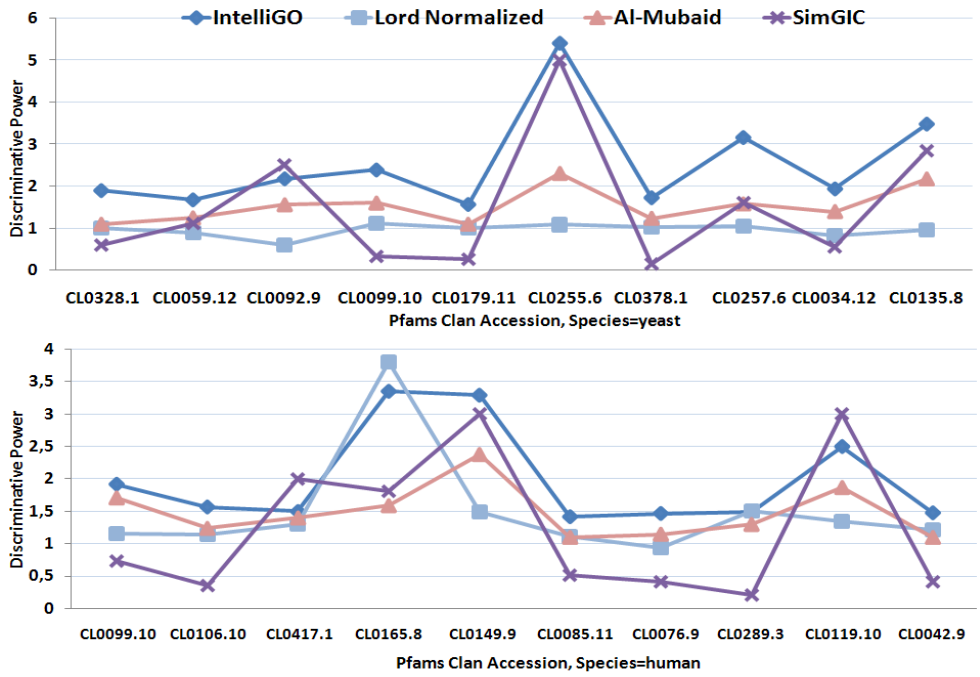


FIGURE 20 – Comparaison des valeurs du pouvoir de discrimination (DP) de quatre mesures de similarité en utilisant les clans Pfam sur les annotations MF. Les valeurs DP obtenus avec *IntelliGO*, Lord-normalisée, Al-Mubaid, et SimGIC sont affichées pour chaque clan pour la levure (Haut) et pour l'humain (Bas).

dans la Table 13. Les valeurs obtenues avec la mesure *IntelliGO* en utilisant les annotations MF en incluant ou en excluant les annotations GO dont le code d'évidence est IEA, sont affichées dans la dernière colonne de la Table 13. Dans le cas où toutes les annotations sont considérées (trois premières lignes), les coefficients de corrélation pour *IntelliGO* varient de 0,40 pour la métrique *SeqSim* à 0,65 pour la métrique *ECC*. La valeur *ECC* obtenue par *IntelliGO* (0,65) est la plus élevée de toutes les valeurs obtenues avec les autres mesures de similarité, y compris la meilleure valeur 0,64 obtenue avec la mesure LB. Pour les métriques *Pfam* et *SeqSim*, les coefficients de corrélation obtenus avec la mesure *IntelliGO* sont moins élevés que les meilleures valeurs obtenues pour cinq et sept autres mesures respectivement. Les meilleures valeurs sont en effet obtenues avec la mesure *SimGIC*, 0,63 pour *Pfam* et 0,71 pour *SeqSim* respectivement. Dans le cas où les annotations de type IEA sont exclues (les trois dernières lignes), la mesure *IntelliGO* génère des coefficients de corrélation en baisse pour les métriques *ECC* et *Pfam*, comme observé pour toutes les autres mesures de similarité, mais en légère hausse pour la métrique *SeqSim* (0,43 au lieu de 0,40). Cette augmentation limitée ou l'absence de diminution est observée avec deux autres mesures (SimUI, JA), alors que beaucoup d'augmentations plus importantes sont enregistrées pour les variantes Maximales des trois mesures de similarité de Resnick, Lord, et Jaccard (RM, LM, JM). De ce fait, il apparait que dans ce type d'évaluation, la mesure *IntelliGO* génère des coefficients de corrélation qui sont intermédiaires entre ceux obtenus avec des mesures de similarité moins performantes (RA, RM, LA, LM, JA, JM) et d'autres mesures plus performantes (SimGIC, SimUI, RB, LB, JB).

Métriques		Mesures de similarité sémantique											
		SimGIC	SimUI	RA	RM	RB	LA	LM	LB	JA	JM	JB	IntelliGO
Tous codes d'évidences	ECC	0.62	0.63	0.39	0.45	0.60	0.42	0.45	0.64	0.34	0.36	0.56	0.65
	Pfam	0.63	0.61	0.44	0.18	0.57	0.44	0.18	0.56	0.33	0.12	0.49	0.48
	SeqSim	0.71	0.59	0.50	0.12	0.66	0.46	0.12	0.60	0.29	0.10	0.54	0.40
Codes d'évidence Non-IEA	ECC	0.58	0.57	0.37	0.47	0.48	0.38	0.51	0.51	0.37	0.46	0.51	0.48
	Pfam	0.58	0.55	0.43	0.44	0.52	0.42	0.42	0.51	0.33	0.34	0.45	0.43
	SeqSim	0.66	0.59	0.46	0.48	0.65	0.41	0.40	0.59	0.31	0.36	0.52	0.43

TABLE 13 – Résultats de l'outil CESSM avec la mesure *IntelliGO*. Les coefficients de corrélation de Pearson sont affichés pour les métriques *ECC* (Enzyme Classification Comparison), *Pfam*, et similarité de séquence (*SeqSim*). Les annotations GO de l'aspect *Molecular Function* sont utilisées en incluant (trois premières lignes) ou en excluant (trois dernières lignes) les termes d'annotations avec les codes d'évidence de type IEA. Les abréviations sont listées comme : SimUI : Union Intersection similarity ; RA : Resnick Average ; RM : Resnick Max ; RB : Resnick Best match ; LA : Lord Average ; LM : Lord Max ; LB : Lord Best match ; JA : Jaccard Average ; JM : Jaccard Max ; JB : Jaccard Best match.

3 Discussion

Dans cette dernière décennie, de plus en plus de mesures de similarité sémantiques et fonctionnelles ont été proposées sur l'ontologie GO. Par conséquent un aspect important de notre étude concernait à la fois la proposition d'une mesure de similarité compétitive, et un outil d'évaluation des mesures de similarité sémantique et fonctionnelle. Dans l'état de l'art, plusieurs approches d'évaluation ont été proposées qui sont à la fois hétérogènes et non reproductibles [PFF+09]. Par exemple, il est généralement supposé que les gènes ayant une similarité de séquence élevée devraient partager des annotations MF similaires. Cette hypothèse a été exploitée par Lord *et al.* pour évaluer leur similarité sémantique en exploitant la corrélation entre les séquences de gènes et leurs annotations dans un ensemble de protéines de l'espèce humaine [LSBG03]. En effet, les auteurs ont trouvé une corrélation entre les annotations de l'aspect MF de GO et les séquences de gènes, et ceci a été confirmé plus tard par Schlicker *et al.* [SDRL06] à l'aide une mesure de similarité différente. Ces derniers auteurs ont aussi testé leur similarité sémantique pour le clustering de familles de protéines à partir de la base de données Pfam en exploitant les annotations MF. Ils ont prouvé que les familles Pfam avec les mêmes fonctions biologiques définissent des clusters de protéines présentant un forte similarité fonctionnelle entre elles. Dans la même perspective, l'outil CESSM a été utilisé dans notre évaluation. Il représente une bonne initiative vers la standardisation des processus d'évaluation des mesures de similarité sémantique. D'autres techniques d'évaluation reposent sur l'hypothèse que les gènes ayant des profils d'expression similaires devraient aussi partager des annotations fonctionnelles similaires. Cette hypothèse a été exploitée par Chabalier *et al.* [CMB07] pour valider leur mesure de similarité sémantique. Ces auteurs ont reconstitué des réseaux de régulations de gènes présentant des similarités très élevées par rapport à leurs annotations BP, et ceci les a conduits à caractériser certains de ces réseaux avec un comportement transcriptionnel particulier, ou à étudier leurs recouvrements avec des pathways de la base KEGG. En utilisant les pathways comme ensembles de gènes préparés par l'expert et représentant une cohésion biologique, des auteurs comme Guo *et al.* [GLS+06] ont démontré que toutes les paires de protéines au sein de pathways KEGG de l'espèce humaine avaient des similarités sémantiques très élevées, plus élevées que celles obtenues avec des ensembles aléatoire de termes d'annotations BP. Wang *et al.* [WDP+07], et Al-Mubaid *et al.* [NAM08b] ont testé quant à eux leurs mesures de similarité sur des gènes de la levure

appartenant à quelques pathways décrits dans la base SGD : *Saccharomyces Genome Database*. Dans la première étude, les auteurs n'ont considéré que les annotations MF pour calculer la similarité sémantique, avec pour objectif de regrouper les gènes ayant des fonctions similaires à l'intérieur d'un seul pathway et en comparant les résultats avec le clustering réalisé avec la mesure de Resnik. Dans la deuxième étude, les valeurs obtenues avec les similarités entre paires de gènes sont calculées en considérant les annotations BP au sein de chaque pathway étudié, et en comparant les résultats avec la même mesure de Resnik.
Dans l'évaluation de notre mesure de similarité *IntelliGO*, nous avons utilisé quatre collections de jeux de données, représentant des pathways KEGG et des clans Pfam. Les expressions *Intra-Set*, *Inter-Set* et *Discriminative Power* ont été introduites afin de réaliser cette évaluation. Nos hypothèses de travail étaient les suivantes.
Premièrement, les gènes appartenant à des pathways KEGG doivent être comparés selon leurs annotations GO dans l'aspect BP car dans un même pathway il existe un ou plusieurs processus biologiques communs.
Deuxièmement, les gènes appartenant à des clans Pfam doivent être comparés selon leurs annotations GO dans l'aspect MF car, dans un même clan, le fait de partager les mêmes domaines Pfam peut refléter des fonctions moléculaires similaires.
Les résultats positifs de l'évaluation de la mesure *IntelliGO* ont permis de confirmer ces hypothèses et nous conduisent à recommander de suivre cette approche pour évaluer d'autres mesures de similarité. Les jeux de données représentent un total de 183+280+118+100=683 gènes respectivement, ce qui est moins grand que le nombre de protéines considérées dans l'outil CESSM (1039). Cependant, les calculs des valeurs *Intra-Set* et *Inter-Set* ont nécessité un total de 67933 paires de comparaisons avec la mesure *IntelliGO* ce qui est largement supérieur au nombre de paires dans l'outil CESSM (13340 paires de protéines). Enfin, les résultats très positifs qui concernent le pouvoir de discrimination de la mesure *IntelliGO* ont ouvert des perspectives pour l'utilisation de cette mesure pour une classification fonctionnelle de gènes, dans le but de regrouper des produits de gènes qui partagent des annotations fonctionnelles sémantiquement similaires.

Chapitre 3

Classification fonctionnelle de gènes avec la mesure de similarité IntelliGO

Sommaire

1 Objectifs

Les algorithmes de "clustering" sont des méthodes de classification non-supervisée très utilisées en fouille de données. Ils utilisent une mesure de similarité ou de distance afin de guider un processus de regroupement d'objets similaires dans des groupes (clusters), et de séparer les objets distants dans des groupes distincts [Mac67]. Comme nous l'avons présenté dans le chapitre 1, il existe une panoplie d'algorithmes de clustering, mais ils obéissent tous au même schéma, c'est à dire : la sélection de variables discriminantes, le choix d'une mesure de similarité ou de distance entre ces variables, un critère de regroupement et la validation des résultats [TK06, Rou87].
Dans le domaine de la génomique fonctionnelle, les applications pour classifier des gènes dans des classes biologiques sont encore peu nombreuses. La classification fonctionnelle a pour objectif de regrouper les gènes selon la similarité de leurs annotations fonctionnelles [BSTP$^+$ce, AS04, NAM08a]. Dans le cas des données d'expression, la classification fonctionnelle permet,

entre autres, de mettre en évidence parmi les gènes qui ont un comportement transcriptionnel particulier (présents dans le même profil) ceux qui partagent des termes d'annotations similaires. Ce principe a été introduit dans le premier chapitre, et vient compléter les techniques d'étude de l'enrichissement des annotations des gènes.
Après avoir démontré le bon comportement de la mesure *IntelliGO* pour des calculs de similarité *Intra-set* et *Inter-set*, j'ai voulu tester cette mesure pour la classification fonctionnelle. Ceci m'a conduit à utiliser des jeux de données contenant des ensembles de référence pour tester et comparer entre elles des mesures de similarité pour la classification fonctionnelle. De plus, j'ai été amené à proposer une méthode pour l'analyse des différences observées entre les clusters fonctionnels et les ensembles de référence.

2 Présentation de l'approche de classification fonctionnelle basée sur la mesure IntelliGO

Le principe général de la classification fonctionnelle est illustré dans la Figure 21. À partir d'une liste de gènes annotés et une mesure de similarité fonctionnelle, une matrice de similarité est calculée. Elle représente les valeurs de similarité sémantique entre toutes les paires possibles des gènes de la liste de départ. Ensuite, un algorithme de clustering exploite la matrice de similarité entre ces gènes pour produire un ensemble de clusters de gènes qui partagent des annotations fonctionnelles similaires.
D'après ce modèle, il paraît clairement que les résultats dépendent fortement d'un côté de la mesure de similarité fonctionnelle et de l'autre, de l'algorithme de clustering utilisé. De plus, et d'après la notion de pouvoir de discrimination introduite dans le chapitre précédent, nous pouvons déjà prédire qu'une mesure de similarité qui a un faible pouvoir de discrimination pourrait générer un clustering de faible qualité.
Notre mesure de similarité *IntelliGO* a été utilisée avec deux techniques de clustering, à savoir les fuzzy c-means (FCM) et le clustering hiérarchique. Les quatre collections d'ensembles de gènes qui ont été utilisées dans la phase d'évaluation de la mesure *IntelliGO*, sont utilisées ici pour réaliser la classification fonctionnelle, c'est à dire 13 pathways de la levure (185 gènes), 13 pathways de l'espèce humaine (280 gènes) pour les annotations BP, 10 clans Pfam de la levure (110 gènes) et 10 autres clans chez l'espèce humaine (100 gènes) pour les annotations MF. Ainsi, les gènes formant tous les ensembles d'une des quatre collections représentent une liste de gènes faisant l'objet d'un clustering. Au total, quatre listes de gènes seront donc utilisées.Les listes de gènes et leur appartenance aux ensembles des jeux de données sont résumées dans l'Annexe A.
Les programmes de clustering utilisés ici sont ceux de la librairie *cluster* disponible dans le package R-Bioconductor, en utilisant les matrices de similarité précédemment calculées. Tous les algorithmes et méthodes d'évaluation du clustering et d'étude de recouvrement qui seront présentés par la suite ont été implémentés en C++, et exécutés sur une machine double CPU 2.40 Ghz avec 4 GO de RAM. Une version en ligne des méthodes de clustering, et les jeux de données sont disponibles dans la plateforme du projet MBI du LORIA : http ://plateforme-mbi.loria.fr/intelligo/.

3 Clustering hiérarchique et visualisation par "heatmap"

En utilisant le clustering hiérarchique (méthode de Ward ascendant) et la visualisation par "heatmap", nous pouvons comparer et évaluer les résultats obtenus avec les mesures de similarité *IntelliGO*, Lord, Al-Mubaid et SimGIC.

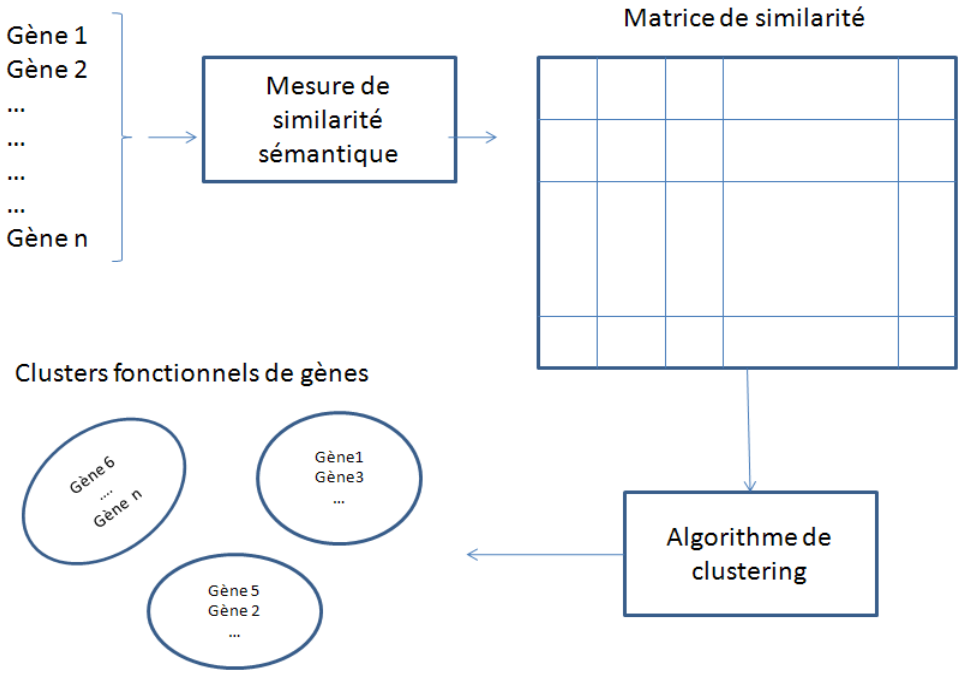


FIGURE 21 – Principe de la classification fonctionnelle. Au départ, une liste de gènes est utilisée avec une mesure de similarité sémantique pour obtenir une matrice de similarité entre toutes les paires de gènes. Ensuite, cette matrice de similarité sera à l'entrée d'un algorithme de clustering pour générer des groupes de gènes fonctionnellement reliés.

Dans la Figure 22, nous avons les résultats qui concernent le clustering des gènes appartenant à 13 pathways KEGG de l'espèce humaine sous forme de quatre heatmaps générées en utilisant respectivement les quatre mesures de similarité sémantique comparées (A) : Lord, (B) : Al-Mubaid, (C) : SimGIC, et (D) : *IntelliGO*. En analysant la heatmap obtenue avec la mesure de Lord (Panneau A), nous pouvons visualiser que les couleurs se distribuent de façon irrégulière sur toute la heatmap, ce qui reflète un regroupement faible de gènes dans des clusters dans la diagonale de la heatmap. En effet, cette technique de visualisation avec un clustering symétrique permet a priori de regrouper le maximum d'éléments similaires dans la diagonale. Les résultats moins performants obtenus avec la mesure de Lord confirment ceux que nous avons déjà obtenus en utilisant l'évaluation par l'indice du pouvoir de discrimination sur les pathways humains (chapitre 2, Figure 19). La situation est pratiquement inversée en considérant les mesures d'Al-Mubaid et SimGIC (Panneaux B et C de la Figure 22). Ces deux heatmaps présentent un nombre limité de clusters de gènes dans la diagonale. La couleur bleu abondante illustre bien la similarité très faible entre les gènes quand on utilise ces deux mesures de similarité. Concernant le Panneau (D) qui correspond à la heatmap d'*IntelliGO*, la distribution des couleurs semble équilibrée pour l'ensemble de gènes. De plus, un nombre supérieur à 10 clusters de tailles différentes pourraient être identifié dans la diagonale de cette heatmap. L'analyse des panneaux (A et B) correspondant aux mesures de similarité de Lord et Al-Mubaid respectivement, révèle que les cellules dans la diagonale ne sont pas toujours de la même couleur, qui devrait être rouge foncé. Ceci signifie qu'il y a des cas où la similarité entre un gène et lui même n'est pas maximale (rouge foncé), ce qui ne respecte pas la condition de la maximalité d'une mesure de similarité [LRB09b], que nous avons définie dans le premier chapitre. Les choses sont différentes avec les mesures SimGIC et

IntelliGO (Panneaux C et D), où la similarité entre un gène et lui-même dans la diagonale est toujours à 1 (rouge foncé).
Ces observations montrent que la heatmap est en elle-même un outil performant pour visualiser le résultat d'un clustering hiérarchique réalisé avec une mesure de similarité fonctionnelle. De plus, il paraît clairement qu'avec un tel type de visualisation, une comparaison préliminaire pourrait être réalisée afin d'estimer la qualité d'une mesure en ce qui concerne son utilisation dans un processus de classification fonctionnelle de gènes.
Des résultats similaires pour les quatre mesures de similarité sont observés dans les Figures (23,24 et 25), où les trois heatmaps qui correspondent à la mesure *IntelliGO* (Panneaux (D) des trois figures) illustrent clairement la performance de cette mesure de similarité fonctionnelle.
J'ai utilisé le clustering hiérarchique et la visualisation par heatmap pour avoir une vision globale sur le comportement d'*IntelliGO* vis-à-vis de la classification fonctionnelle. Puisque cette méthode de classification ne permet pas d'affecter un gène à plusieurs clusters à la fois, j'ai par la suite décidé d'utiliser une approche de "clustering flou". Seule la mesure de similarité *IntelliGO* sera considérée pour la suite, puisqu'elle a montré une bonne performance pour la classification fonctionnelle.

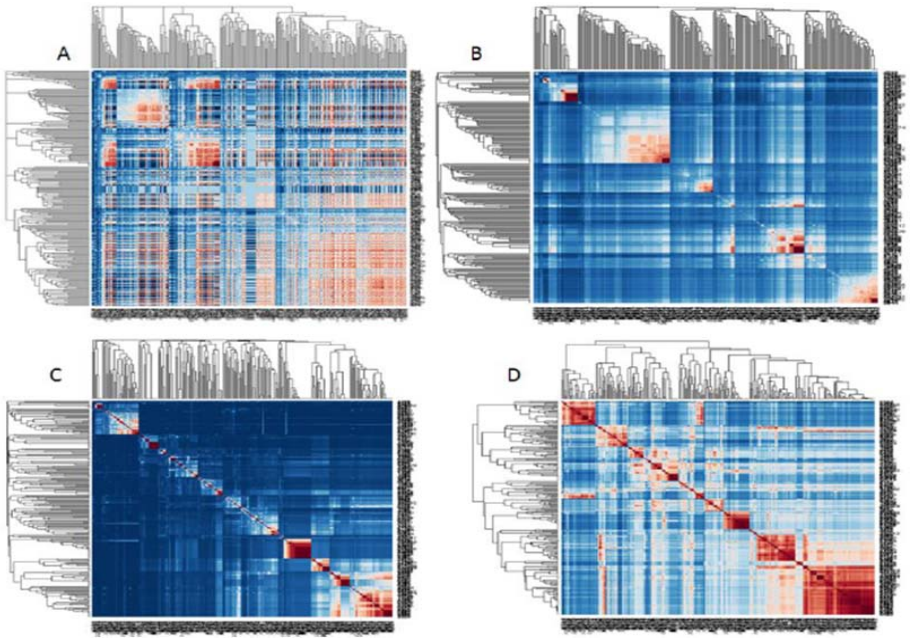


FIGURE 22 – Clustering hiérarchique et visualisation par heatmaps, générées en utilisant des matrices de similarité obtenues avec les similarités de (A) : Lord, (B) : Al-Mubaid, (C) : SimGIC, et (D) : *IntelliGO*. Les gènes appartiennent aux 13 pathways KEGG de l'espèce humaine décrits au Chapitre 2 et dans l'Annexe 1. Les couleurs varient du rouge foncé pour les paires de gènes très similaires, vers le bleu foncé pour les gènes les moins similaires.

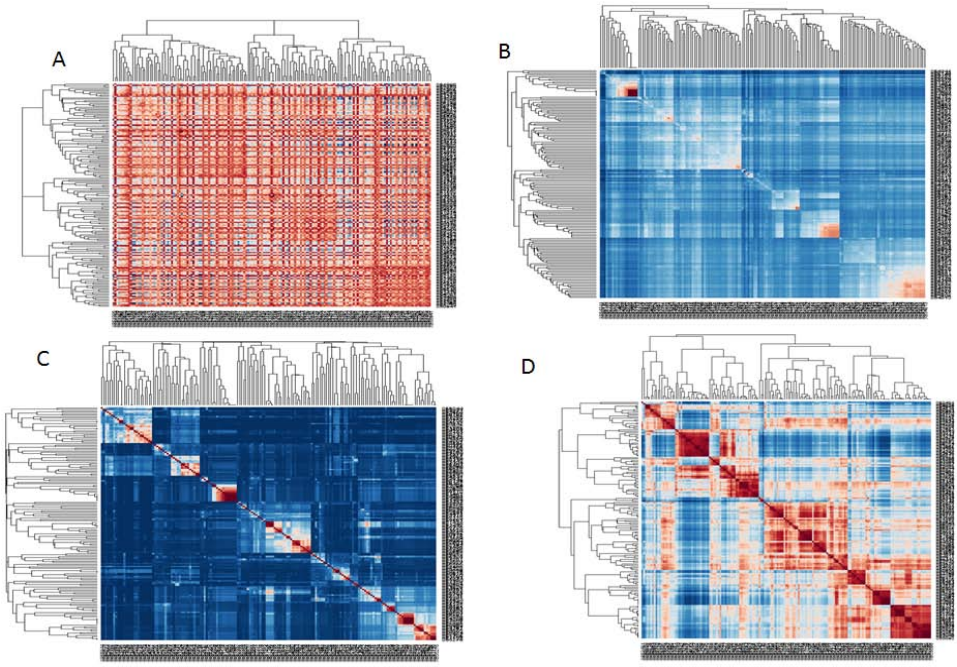


FIGURE 23 – Clustering hiérarchique et visualisation par heatmaps, générées en utilisant des matrices de similarité obtenues avec les similarités de (A) : Lord, (B) : Al-Mubaid, (C) : SimGIC, et (D) : *IntelliGO*. Les gènes appartiennent aux 13 pathways KEGG de l'espèce levure décrits au Chapitre 2 et dans l'Annexe 1. Les couleurs varient du rouge foncé pour les paires de gènes très similaires, vers le bleu foncé pour les gènes les moins similaires.

4 Classification fonctionnelle floue de gènes avec la mesure IntelliGO et comparaison avec l'outil DAVID

4.1 Généralités

A ce stade de notre étude, j'ai décidé d'utiliser le clustering flou (FCM) qui vient compléter le clustering hiérarchique en permettant un chevauchement entre clusters. Rappelons que l'algorithme FCM a été utilisé dans de nombreuses applications dans le domaine bioinformatique. Par exemple, Eisen *et al.* ont utilisé cette approche pour classifier les données d'expression dans des profils d'expression [GE02]. L'outil "DAVID functional classification" représente aussi un bon exemple où la méthode de classification par FCM a été utilisée pour regrouper les gènes ayant des annotations fonctionnelles similaires, en utilisant la statistique Kappa [HST+07]. L'intérêt d'utiliser la mesure de similarité sémantique *IntelliGO* et le clustering flou vient du fait que cette mesure s'est montrée performante concernant l'évaluation par des indices intrinsèques (ex : $Intra - Set$) et extrinsèques (ex : Pouvoir de discrimination) que nous avons présentés dans le chapitre précédent. Ces critères d'évaluation sont à rapprocher de plusieurs indices utilisés pour évaluer les méthodes de clustering [KLL04, WY05]. J'ai proposé une autre approche d'évaluation du clustering flou basé sur *IntelliGO* en utilisant les ensembles de référence qui constituent les collections de gènes que nous avons utilisées. Cette approche d'évaluation du clustering a débouché vers une technique d'étude de recouvrement entre clusters fonctionnels de gènes et les ensembles de référence.
Les résultats du clustering générés avec la mesure de similarité *IntelliGO* sont comparés avec

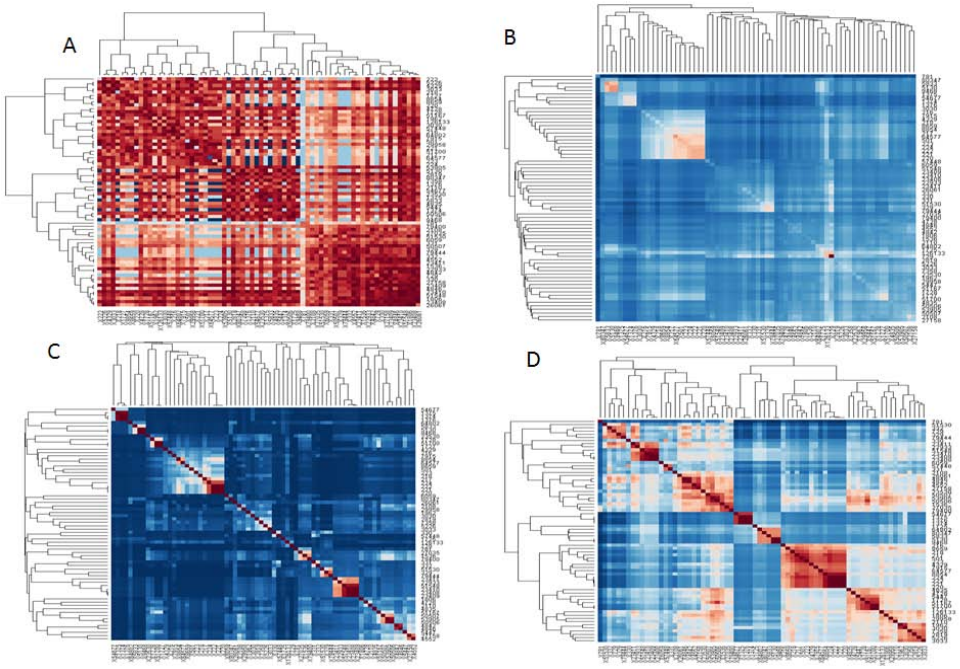


FIGURE 24 – Clustering hiérarchique et visualisation par heatmaps, généré en utilisant des matrices de similarité obtenu avec les similarités de (A) : Lord, (B) : Al-Mubaid, (C) : SimGIC, et (D) : *IntelliGO*. Les gènes appartiennent aux 10 clans Pfam de l'espèce humaine décrits au Chapitre 2 et dans l'Annexe 1. Les couleurs varient du rouge foncé pour les paires de gènes très similaires, vers le bleu foncé pour les gènes les moins similaires.

ceux obtenus avec l'outil "DAVID functional classification tool" sur la base du calcul d'un F-score pour chaque clustering. J'ai utilisé comme dans le cas du clustering hiérarchique, les matrices de similarité *IntelliGO* obtenues avec les listes de gènes des quatre collections d'ensembles de gènes, pour servir comme entrées à l'algorithme FCM. La méthode d'évaluation du clustering proposée permet à la fois d'estimer le meilleur nombre de clusters à générer, et de comparer deux résultats de clustering. Le détail de l'approche ainsi que les différents algorithmes que nous avons proposés, sont présentés avec des exemples dans les sous-sections suivantes.

4.2 Technique d'évaluation d'un clustering avec la mesure F-score et des ensembles de référence

Dans le cas où des ensembles de référence de gènes sont disponibles, la méthode du F-score permet d'exprimer le degré de recouvrement entre les ensembles de référence et les clusters produits [vR79]. Cette méthode est souvent utilisée en recherche d'information, et fournit une estimation quantitative de l'efficacité de l'appariement en fonction de la précision (spécificité) et du rappel (sensibilité). La précision (en anglais precision), représente le nombre d'objets pertinents retournés sur le nombre total d'objets sélectionnés. Le rappel (en anglais recall), représente le nombre d'objets pertinents retournés sur le nombre total d'objets pertinents. La formule du F-score est donnée par la formule suivante :

$$F - score = \frac{2 * Precision * Recall}{Precision + Recall}. \quad (1)$$

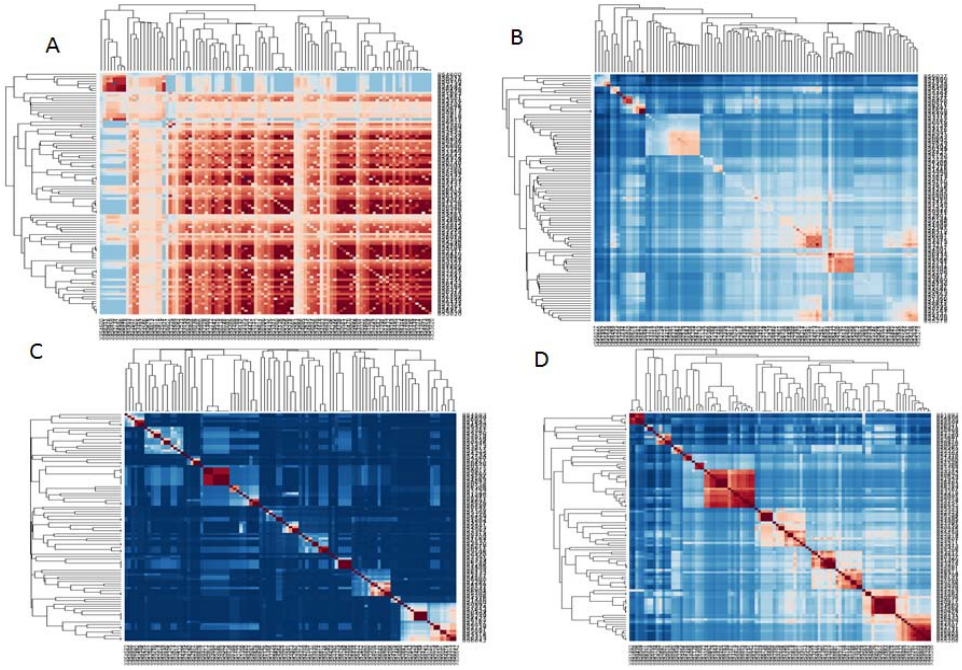


FIGURE 25 – Clustering hiérarchique et visualisation par heatmaps, généré en utilisant des matrices de similarité obtenu avec les similarités de (A) : Lord, (B) : Al-Mubaid, (C) : SimGIC, et (D) : *IntelliGO*. Les gènes appartiennent aux 10 clans Pfam de l'espèce levure décrits au Chapitre 2 et dans l'Annexe 1. Les couleurs varient du rouge foncé pour les paires de gènes très similaires, vers le bleu foncé pour les gènes les moins similaires.

L'algorithme (2) décrit une méthode d'optimisation d'un clustering en utilisant des ensembles de référence et la méthode F-score. Tout d'abord, nous considérons Σ=$\{R_1, R_2, .., R_p\}$ une collection d'ensembles de référence de gènes. Σ pourrait être une des collections présentées précédemment.
Comme dans le cas du clustering hiérarchique, tous les gènes dans la collection $\bigcup R_i$ sont utilisés pour obtenir une matrice de similarité. Celle-ci sera utilisée pour générer K clusters Φ=$\{C_1, C_2, ..., C_K\}$. Nous varions le nombre K de clusters à produire par l'algorithme de clustering dans un intervalle p∈[n_1,n_2] au voisinage du nombre p d'ensembles de référence, en supposant qu'un bon clustering devrait produire un nombre optimal de clusters proches du nombre d'ensembles de référence [BSTP$^+$ce].
Pour chaque ensemble de référence R_i, on parcourt chaque cluster généré C_j pour calculer les valeurs : Precision(R_i,C_j)= $|R_i \cap C_j|/|C_j|$, Recall(R_i,C_j)= $|R_i \cap C_j|/|R_i|$ et F-score(R_i, C_j)= $\frac{2*Precision(R_i,C_j)*Recall(R_i,C_j)}{Precision(R_i,C_j)+Recall(R_i,C_j)}$. Ensuite, pour tous les F-scores calculés pour chaque paires (R_i, C_j), une valeur F-score maximale $\widehat{F-score}(R_i) = Max_{C_j \in \Phi}(F - score(R_i, C_j))$ est calculée. La moyenne des F-scores maximums est alors calculée pour donner la valeur du F-score global associé au résultat du clustering étudié. Pour terminer, la valeur $\widehat{K}$ donnant le F-score global maximum, est retournée comme meilleur nombre K de clusters générés.
Concernant le clustering flou basé sur *IntelliGO* en utilisant les collections 1 et 2, nous avons fait varier le nombre de clusters à générer K entre 11 et 17 avec un pas de 1 puisque ces deux collections comportent 13 pathways chacune pour la levure et l'humain.Pour les collections 3 et 4, les valeurs de K varient entre 8 et 14 avec un pas de 1, puisque ces deux collections comportent 10 clans Pfam chacune pour les deux espèces.Ainsi, pour chaque valeur de K, un score

Algorithm 2 Optimisation du Clustering avec les ensembles de références et la mesure F-score.

Require: Σ=$\{R_1, R_2, .., R_p\}$: une collection d'ensembles de référence, deux entiers(n_1,n_2) avec p∈[n_1,n_2], une matrice de similarité entre tous les éléments de Σ.

Ensure: Le nombre optimal de clusters à générés $\widehat{K}$, et une valeur $Global\ \widehat{F-score}(\widehat{K})$.

1: **for** chaque $K \in [n_1,n_2]$ **do**
2: Générer K clusters Φ=$\{C_1, C_2, ..., C_K\}$, en utilisant tout les éléments dans $\bigcup R_i$
3: **for** chaque ensemble de référence $R_i \in \Sigma$ **do**
4: **for** chaque cluster $C_j \in \Phi$ **do**
5: Precision(R_i,C_j)=$|R_i \cap C_j|/|C_j|$
6: Recall(R_i,C_j)= $|R_i \cap C_j|/|R_i|$
7: $F-score(R_i,C_j)=\frac{2*Precision(R_i,C_j)*Recall(R_i,C_j)}{Precision(R_i,C_j)+Recall(R_i,C_j)}$
8: **end for**
9: $\widehat{F-score}(R_i) = Max_{C_j \in \Phi}(F-score(R_i,C_j))$
10: **end for**
11: $Global\ F-score(K) = \frac{\sum_{i=1}^{p}(|R_i|*\widehat{F-score}(R_i))}{\sum_{i=1}^{p}|R_i|}$
12: **end for**
13: $Global\ \widehat{F-score}(\widehat{K}) = Max_K Global\ F-score(K)$
14: **return** $\widehat{K}$, $Global\ \widehat{F-score}(\widehat{K})$.

$Global\ F-score(K)$ est calculé.

Parallèlement, le clustering fonctionnel avec l'outil DAVID est réalisé en considérant les mêmes jeux de données. Nous avons fait varier le seuil de similarité *Kappa* entre $0,3$ et $0,7$ avec un pas de $0,1$ dans le but d'obtenir différents nombres de clusters, puisque DAVID ne permet pas de spécifier le nombre de clusters *a priori*. L'outil de classification DAVID prend en entrée ce seuil de similarité Kappa, et ne garde parmi les gènes en entrée, que les paires dont la similarité Kappa est supérieure ou égale à ce seuil. Par conséquent, si on choisit un seuil faible, beaucoup de gènes sont clusterisés mais le résultat est moins performant. Tandis que si on fixe un seuil Kappa vers une valeur élevée, cela signifie que le clustering sera plus performant mais l'outil exclut les paires de gènes à partir de la liste d'entrée, ayant la similarité Kappa inférieure au seuil prédéterminé. Cette perte de gènes qui est une particularité de l'outil DAVID est inévitable car il n'y a pas d'autre moyen de faire varier le nombre de clusters à produire. Elle doit être prise en compte pendant l'interprétation des résultats de la comparaison. Comme dans le cas précédent, les K clusters sont appariés avec les ensembles de référence en entrée, et une valeur $Global\ F-score(K)$ est calculée.

Les résultats obtenus avec *IntelliGO* et DAVID sont récapitulés dans la table 14. Concernant la première collection d'ensembles de gènes (13 pathways KEGG humain), en utilisant notre mesure de similarité, nous pouvons constater que toutes les valeurs de F-score sont supérieures à $0,5$, avec un maximum de $0,62$ pour $K = 14$. Cela signifie que les gènes des 13 pathways humains se regroupent bien dans 14 clusters fonctionnels. Ceci pourrait être expliqué par le fait qu'un pathway de la base KEGG pourrait correspondre à plus d'un processus biologique et/ou que la classification fonctionnelle *IntelliGO* a regroupé des gènes de divers pathways partageant des annotations BP sémantiquement similaires en un pathway supplémentaire.

Avec la classification DAVID, le F-score maximum est atteint à $0,67$ pour *Kappa* = 0,3, donnant 10 clusters fonctionnels. Cependant, ce résultat est obtenu avec un seuil Kappa le plus faible. En effet, avec un seuil très élevé (0,7) le nombre de gènes exclus du clustering est très important, mettant en évidence clairement les limites de cet outil. Des résultats très similaires sont obtenus

avec les collections 2, 3 et 4 comme illustré dans la Table 14.

4.3 Discussion

En résumé, ces résultats montrent que la classification fonctionnelle *IntelliGO* est une alternative compétitive à l'outil de classification DAVID. En effet, d'une part, le nombre optimal de clusters *IntelliGO* est toujours proche du nombre d'ensembles de référence de la collection. D'autre part, la classification *IntelliGO* n'exclut aucun gène pendant le processus, contrairement à l'outil DAVID avec lequel les résultats obtenus sont de faible qualité puisque les meilleurs F-score correspondent aux faibles valeurs Kappa (0.3). Suite à ces résultats, j'ai proposé de mettre en ligne notre outil de classification *IntelliGO* pour qu'il puisse être testé par d'autres utilisateurs et sur d'autres jeux de données (http ://plateforme-mbi.loria.fr/intelligo). L'analyse préliminaire du nombre d'accès à ces outils nous a fourni une idée sur le grand intérêt qu'à suscité les utilisateurs de différentes nationalités.

5 Analyse de recouvrement entre ensembles de référence et clusters fonctionnels

5.1 Appariement cluster-ensemble de référence

Afin de raffiner notre comparaison, nous avons décidé d'analyser de près les résultats de recouvrement entre les ensembles de référence et les clusters obtenus avec un nombre optimal K de clusters. Nous avons utilisé l'algorithme 3 afin d'extraire le cluster se trouvant en haut de la liste des clusters qui se recouvrent avec un ensemble de référence, quand on les trie par valeur décroissante du F-score. Dans un cas avec les pathways KEGG et dans trois cas avec les clans Pfam, un recouvrement parfait est observé avec un $F - score$=1. Cela veut dire que la précision et le rappel de recouvrement sont aussi à 1. De tels recouvrements parfaits sont observés uniquement avec la classification *IntelliGO*. Concernant la classification DAVID, des situations d'ambiguïté de recouvrement sont observées dans huit cas pour les pathways KEGG et deux cas pour les clans Pfam, où le même cluster est affecté à deux ensembles de référence distincts avec des valeurs F-score comparables.
Cette analyse confirme que pour les quatre collections utilisées ici, le clustering flou basé sur la mesure *IntelliGO* produit un nombre optimal de clusters K qui recouvrent au mieux les ensembles de référence, de façon plus performante que le clustering fonctionnel DAVID, sans exclure le moindre gène de l'analyse. Cette analyse constitue la première étape de l'approche d'analyse de recouvrement ("overlap analysis"). Dans ce qui suit, nous exploitons la différence ensembliste dans les cas de meilleur recouvrement entre les paires d'ensembles de référence et de clusters, pour des valeurs F-score inférieures à 1.

5.2 Analyse des différences ensemblistes dans les paires les mieux appariées : Découverte d'informations manquantes.

Pour une paire R,C d'ensemble de référence R et de cluster C appariés selon l'algorithme (3), l'intersection $R \cap C$ est supposée refléter un contenu homogène de gènes présents à la fois dans un ensemble de référence et dans un cluster fonctionnel. Alternativement, les deux différences d'ensembles $C \backslash R$ et $R \backslash C$ pourraient être considérées comme sources de connaissances pour la

	A. IntelliGO		B. DAVID classification tool			
Collection	K	Global F-score	Seuil Kappa	K	Global F-score	% exclusion
1	11	0.59	**0.3**	**10**	**0.67**	20.7
(280 gènes)	12	0.61	0.4	11	0.63	31.4
	13	0.61	0.5	14	0.66	38.2
	14	**0.62**	0.6	11	0.41	75.9
	15	0.56	0.7	8	0.31	68.2
	16	0.55				
	17	0.54				
2	11	0.59	**0.3**	**9**	**0.68**	17.8
(185 gènes)	12	0.62	0.4	8	0.65	31.4
	13	0.64	0.5	9	0.55	43.2.
	14	**0.67**	0.6	6	0.39	56.8
	15	0.66	0.7	7	0.20	69.2
	16	0.62				
	17	0.62				
	8	0.70	**0.3**	**11**	**0.64**	27.0
3	9	0.64	0.4	11	0.51	52.0
(100 gènes)	10	0.68	0.5	8	0.31	66.0
	11	**0.75**	0.6	3	0.09	93.0
	12	0.66	0.7	2	0.01	96.0
	13	0.66				
	14	0.64				
	8	0.79	**0.3**	**10**	**0.70**	40.7
4	9	0.77	0.4	9	0.47	61.0
(118 gènes)	10	0.78	0.5	9	0.39	69.5
	11	**0.82**	0.6	5	0.28	84.7
	12	0.78	0.7	3	0.21	91.5
	13	0.78				
	14	0.71				

TABLE 14 – Variation du F-score Global en faisant en varier (A) le nombre K de clusters flous générés en utilisant *IntelliGO* et le programme FCM, et (B) le seuil Kappa avec l'outil de classification DAVID. Dans B est indiqué le pourcentage de gènes exclus du clustering dans la colonne (% exclusion). Les résultats sont affichés pour les quatre collections de références (rappel du nombre de gènes total entre parenthèses). Le nombre K optimal et le F-score maximum correspondant sont indiqués en gras.

Algorithm 3 Affectation d'un cluster à un ensemble de référence selon la valeur $F - score$.

Require: Σ=$\{R_1, R_2, .., R_p\}$: une collection d'ensemble de référence, Φ_K=$\{C_1, C_2, ..., C_K\}$: une collection de clusters, $\forall(i,j)$ | 1≤i≤p, 1≤j≤K, $F - score(R_i, C_j)$ (voir Algorithme 2).
Ensure: Un cluster assigné à chaque ensemble de référence.
1: **for** chaque ensemble de référence $R_i \in \Sigma$ **do**
2: $List_i$⟵(C_j,$F - score(R_i,C_j)$) : Une liste de clusters C_j classée par ordre décroissant selon le F-score(R_i,C_j)
3: C⟵$List_i[1]$: Un cluster qui est dans le haut de la $List_i$
4: **print** (R_i,C)
5: **end for**

découverte d'informations manquantes (*missing information approach*) [BSTND].
En considérant la différence $C\backslash R$ qui contient les gènes qui sont similaires aux gènes de l'ensemble de référence mais ne font pas partie de ses membres, cette différence pourrait être exploitée par un expert pour vérifier s'il y a des gènes de $C\backslash R$ qui pourraient être affectés à R (Algorithme 4). Concernant la deuxième différence d'ensemble $R\backslash C$, elle est supposée contenir les gènes qui sont membres d'un ensemble de référence mais qui n'ont pas été regroupés au sein d'un cluster avec les autres membres de l'ensemble de référence qui partagent des annotations similaires. Les annotations fonctionnelles des gènes de $R\backslash C$ pourraient être scrutées par un expert dans le but de vérifier si l'annotation de ces gènes pourraient être complétée par l'un ou l'autre des termes significatifs de C (Algorithme 5). Pour les deux algorithmes, un poids de suggestion a été proposé pour faciliter la tâche de l'expert.
Ces deux algorithmes permettent donc la découverte d'informations manquantes, après l'analyse des différences ensemblistes $C\backslash R$ et $R\backslash C$ pour les paires fortement appariées. J'ai implémenté ces deux algorithmes, en les exécutant pour la classification *IntelliGO* et DAVID, en utilisant les quatre collections d'ensembles de référence. Comme dans le cas précédent, nous considérons ici pour les deux méthodes de clustering, le nombre optimal K de clusters à générer obtenus avec les algorithmes précédents (2,3). La valeur δ_{P_value} qui correspond dans les algorithmes (4,5) au seuil de significativité de l'enrichissement des termes GO, a été fixée à 10^{-4} qui est une valeur classiquement choisie dans de telles analyses. Les suggestions d'enrichissement retournées par chaque algorithme, sont de l'ordre d'une centaine, et ont été analysées par l'expert biologiste. Nous mettons en évidence ici quatre suggestions importantes, trois obtenues avec l'analyse de la classification *IntelliGO* et une obtenue avec la classification DAVID.
Seulement un cas résulte de l'analyse $C\backslash R$ avec l'algorithme (4) : le gène humain *STON1* a été suggéré appartenir au pathway *hsa*04130 (SNARE interactions in vesicular transport) sur la base de son annotation BP "intracellular protein transport". Cette suggestion n'a pas pu être confirmée par l'expert mais a révélé que le gène *STON1* a été mal classé dans le pathway *hsa*03022 (basal transcription factors).
Trois autres cas concernent des annotations manquantes (missing annotations) révélées par l'analyse de $R\backslash C$ selon l'algorithme 5. Par exemple, le terme GO "methionine biosynthetic process" pourrait être ajouté aux annotations du gène ARO8 de la levure parce que ce terme est significativement enrichi dans la paire la mieux appariée entre le cluster C_9 et le pathway de la levure *sce*00300 (Lysine metabolism) qui contient le gène ARO8. L'expert a validé cette suggestion car le gène ARO8 est aussi membre du pathway cysteine and methionine metabolism *sce*00270.
Un autre exemple est trouvé dans l'analyse $R\backslash C$ avec la troisième collection d'ensembles (clans Pfam humain). L'algorithme 5 propose l'ajout du terme GO de l'aspect MF "enzyme binding" à trois gènes membres du clan Pfam humain *CL*0417.1 (BIR-like), qui sont *BIRC3*, *BIRC4* and

BIRC6. L'expert a aussi validé cette proposition car les trois gènes ont la propriété de lier la caspase qui est une enzyme.
Le troisième exemple a été obtenu en analysant $R\backslash C$ avec la classification DAVID, et a été validé par l'expert. Le terme "AMP binding" de l'aspect MF, est proposé pour faire partie des annotations du gène de la levure YOR093C (NCBI Gene Id=854260), qui est membre du clan $CL0378$ (Ac-CoA-synthase). Effectivement, l'expert a pu vérifier que ce gène contient un domaine de liaison à l'AMP.

Algorithm 4 Enrichissement d'ensemble par analyse de $C\backslash R$.

Require: (R,C) : une paire mieux appariée (ensemble de référence, cluster). δ_{P_value} : un seuil de p-value.
Ensure: Liste de gènes suggérés d'un cluster C_j à ajouter à un ensemble de référence R_i.

```
Calculer Enrichment(R)={terms ti | P_value(ti) ≤ δ_P_value}
if Enrichment(R)≠∅ then
  for chaque gène g∈C\R do
    suggested=0
    Extraire Annotation(g)={t1, t2, ..., tn | ti annotates g}
    for chaque tk∈Annotation(g) do
      if (tk∈Enrichment(Ri)) then
        suggested++
      end if
    end for
    Weight_of_Suggestion=suggested / |Enrichment(Ri)|
    if suggested≠0 then
      print (g , Weight_of_Suggestion)
    end if
  end for
end if
```

6 Discussion

Les exemples montrent que l'approche proposée ici pourrait être appliquée pour découvrir des gènes mal classés ou des annotations manquantes de gènes. Les résultats préliminaires sont en effet prometteurs et la méthode pourrait être utilisée dans des processus d'amélioration du contenu des bases de données biologiques ("biological database curation") ou pour l'annotation des génomes [P08]. Des expériences dédiées impliquant des ensembles de données appropriés devraient être conçues à cette fin.
Nous avons proposé la mesure de similarité sémantique *IntelliGO*, avec comme principal objectif la classification fonctionnelle de gènes. Les résultats que nous avons présentés illustrent bien la performance de la classification *IntelliGO* par rapport au l'outil de base DAVID souvent utilisé pour l'analyse fonctionnelle de gènes. La robustesse de la classification est prouvée grâce aux différentes évaluations que nous avons menées en utilisant différentes collections de jeux de données en variant l'espèce et l'aspect des annotations de gènes. Les évaluations concernaient à la fois les critères de qualité intrinsèques aux clusters de gènes (Discriminative Power) et extrinsèque en utilisant la méthode de recouvrement avec les ensembles de référence et le F-score.

Algorithm 5 Enrichissement d'annotations de gènes par analyse $R\backslash C$.

Require: (R,C) : une paire mieux appariée (ensemble de référence, cluster). δ_{P_value} : un seuil de p-value.

Ensure: Liste de termes d'annotation à ajouter à des gènes de l'ensemble de référence R_i.

Calculer Enrichment(C)=$\{terms\ ti \mid P_{value}(ti) \leq \delta_{P\_value}\}$
for chaque gène $g \in R\backslash C$ **do**
  Extraire Annotation(g)=$\{t_1, t_2, ..., t_n \mid t_i\ annotates\ g\}$
end for
if $R\backslash C \neq \emptyset$ **then**
  for chaque $t_k \in Enrichment(C)$ **do**
    for chaque gène $g \in R\backslash C$ **do**
      if ($t_k \notin Annotation(g)$) **then**
        $Suggested\_list_k \longleftarrow$ g
      end if
    end for
    $Weight\_of\_Suggestion(t_k) = \frac{|Suggested\_list_k|}{|R\backslash C|}$
    if $Suggested\_list_k \neq \emptyset$ **then**
      print (t_k,$Suggested\_list_k$,
      $Weight\_of\_Suggestion(t_k)$)
    end if
  end for
end if

La pertinence des résultats biologiques a été validée par l'expert et a contribué à la proposition de nouveaux éléments pour enrichir les ensembles de gènes disponibles dans les bases de données biologiques. Nous rappelons ici que l'approche de la classification *IntelliGO* a été proposée pour interpréter les données d'expression, et pour compléter les outils d'étude d'enrichissement qui existent déjà.

Tout ce que nous avons présenté jusqu'à maintenant concernait les données symboliques, c'est à dire les annotations fonctionnelles de gènes. Nous avons présenté dans le premier chapitre l'allure générale des données transcriptomiques (Table 3, en section 3). Nous allons dans le chapitre suivant, présenter une nouvelle approche qui concerne la prise en compte des données numériques que sont les niveaux d'expression de gènes.

Chapitre 4

Extraction de profils flous d'expression différentielle

Sommaire

1 Profils flous d'expression différentielle : Fuzzy DEP

1.1 Motivations

Dans les approches classiques d'extraction de profils d'expression, aucune information a priori sur la nature des situations biologiques n'est prise en considération. Or, l'interprétation du regroupement des gènes dans des profils d'expression est souvent guidée par une hypothèse a priori exprimant l'objectif de l'étude. C'est le cas en particulier lorsque des relations existent entre les situations étudiées. Ces relations peuvent être temporelles lorsqu'il s'agit des différents points d'une cinétique ou physiologiques lorsqu'il s'agit d'un même type de tissu pris dans des conditions physiologiques différentes. Dans ce cas, ce sont les différences entre les situations prises deux à deux qui intéressent le chercheur et il serait intéressant de regrouper ensemble les gènes qui présentent les mêmes variations entre deux situations indépendamment du niveau d'expression qu'ils présentent chacun dans l'une et l'autre des situations. Cette constatation nous a conduit à introduire ici la notion de définition a priori de *profils d'expression différentielle* (DEP : Differential

Expression Profiles). Construits à partir de paires de situations choisies par l'utilisateur, ces profils peuvent être considérés comme des combinaisons, pour chaque paire de situations, d*'ensembles d'expression différentielle* (EED) destinés à regrouper les gènes présentant les mêmes variations (sur-expression, sous-expression ou même expression) entre deux situations. L'appartenance d'un gène à tel ou tel profil commence donc par l'étude de son appartenance à tel ou tel ensemble d'expression différentielle pour chaque paire de situations. Nous utilisons ici une modélisation floue [Kos99, TN06] des variations d'expression entre deux situations afin de tenir compte du bruit des données et de donner la possibilité à un gène d'appartenir à plus d'un profil.

1.2 Présentation du contexte biologique

Les données d'expression de gènes peuvent être représentées par une matrice M : Gènes $\times$ Situations dans laquelle $M_{i,j}$ représente la mesure de l'expression du gène g_i dans la situation S_j.
Nous avons récupéré grâce à une collaboration avec des biologistes de l'IGBMC de Strasbourg, un jeu de données transcriptomiques relatif au cancer colo-rectal. Une analyse sur une puce à ADN de type Affymetrix HGU133+ [12], contenant environ 51227 sondes (probes set) a été réalisée (expérience du 8 Février 2008), et trois types de tissu ont été utilisés : tissu normal avec 48 réplicats, tissu cancéreux avec 56 réplicats, et un tissu de lignée cellulaire avec 35 réplicats. Une étape de filtrage a permis de garder 36296 sondes représentant des gènes dont l'expression est significative soit 70% des sondes initiales. Afin de tenir compte de la variabilité biologique, la mesure d'expression d'un gène dans une situation résulte de l'analyse répétée 3 fois de plusieurs dizaines d'échantillons de même type biologique. On calcule trois valeurs moyennes pour chaque type de tissu pour produire trois situations. Ensuite, une analyse d'identification de gènes différentiellement exprimés a été faite par des approches statistiques à l'IGBMC, ce qui a conduit à la création d'une liste de 222 gènes différentiellement exprimés.
Nous nommons ici les trois situations considérées et qui correspondent aux échantillons (i) de tissu colo-rectal sain, (ii) de tumeur colo-rectale, (iii) de lignée cellulaire cancéreuse en culture par : (S_1), (S_2) et (S_3) respectivement. Les données d'expression correspondantes sont représentées dans la matrice d'expression M comme illustré dans la Table 15, où les valeurs d'expression ν_{s1}, ν_{s2}, ν_{s3} fournies pour chaque gène du tableau sont alors des moyennes des valeurs observées pour chaque échantillon et chaque réplicat. Le sous-ensemble de 222 gènes qui a été sélectionné à partir du jeu initial de gènes analysés sur la puce, présente un changement significatif des valeurs d'expression entre les situations S_1 et S_2. De fait, ce qui intéresse le biologiste ici ce sont essentiellement les gènes dont l'expression varie (on dit alors qu'ils sont dé-régulés) entre le tissu sain (S_1) et le tissu cancéreux (S_2). Le comportement de ces gènes dans les lignées cancéreuses (S_3) représente une sorte de référence pour le niveau d'expression des gènes dé-régulés de façon générale dans la plupart des cancers.

1.3 Définition des profils d'expression différentielle

Dans l'étude présentée ici, il apparaît que les situations considérées dans notre jeu de données, sont reliées physiologiquement. Nous introduisons ici la notion de *profil d'expression différentielle* qui prend en compte pour chaque paire de situation considérée, la variation d'expression entre ces deux situations. Dans le cas présent avec trois situations, cela nous conduit à envisager 3 paires de situations : (S_3, S_1) (S_2,S_1) et (S_3,S_2). Respectivement pour n situations, il faudrait envisager $n(n-1)/2$ paires de situations. Déterminer *un profil d'expression différentielle* pour

12. www.affymetrix.com/products_services/arrays/specific/hgu133plus.affx

Gène	Situation 1 (Tissu normal) : S_1	Situation 2 (Tissu cancéreux) : S_2	Situation 3 (Lignée cellulaire) : S_3
KIAA1199	33,6	827,87	735,75
FOXQ1	65,36	1240,21	2631,71
PSAT1	89,03	1019,0	3025,66
CLDN1	12,15	119,9	78,5
SLC6A6	56,6	551,1	568,6
....	...	...	...
Gène g	ν_{s1}	ν_{s2}	ν_{s3}
....	...	...	...
PSAT1	113,1	407,1	1258,0

TABLE 15 – Exemple de matrice d'expression M d'une liste de 222 gènes relatifs aux cancers colorectaux. Cette matrice d'expression est utilisée dans l'étape d'extraction de profils d'expression différentielle.

un gène g revient alors à placer ce gène, pour chaque paire de situations (S_i, S_j), soit dans l'ensemble des gènes sur-exprimés en S_i par rapport à S_j (que nous noterons $\text{Over}_{i,j}$) soit dans l'ensemble des gènes sous-exprimés en S_i par rapport à S_j (que nous noterons $\text{Under}_{i,j}$), soit encore dans l'ensemble des gènes qui sont autant exprimés en S_i qu'en S_j (que nous noterons $\text{Iso}_{i,j}$).

Définition: Les ensembles $\text{Over}_{i,j}$, $\text{Under}_{i,j}$, $\text{Iso}_{i,j}$ sont notés *ensembles d'expression différentielle*, permettant de représenter la différence d'expression des gènes entre les situations i et j.

Définition: La représentation formelle d'un *profil d'expression différentielle* PED est donc un *k-uplet* d'ensembles d'expression différentielle, k étant le nombre de paires de situations considérées.

Par exemple, dans notre contexte applicatif, le profil ($\text{Over}_{3,1}$, $\text{Over}_{2,1}$, $\text{Under}_{3,2}$), est un profil d'expression différentielle, dans lequel les gènes sont sur-exprimés dans la lignée cancéreuse et dans la tumeur colo-rectale par rapport au tissu sain, et sous-exprimés dans la lignée cellulaire par rapport à la tumeur colo-rectale [BDST+09].

1.4 Modélisation floue de l'appartenance à un ensemble d'expression différentielle

La prise en compte de la variabilité biologique et du bruit, inhérent aux procédés de mesure nous conduit à proposer une modélisation *floue* de la fonction d'appartenance à tel ou tel ensemble d'expression différentielle. En effet pour certaines valeurs limites, il n'est pas possible de décider si le gène est sur-exprimé (respectivement sous-exprimé) ou exprimé au même niveau dans deux situations.
Un premier choix dans la modélisation concerne le calcul de la variation d'expression entre deux situations S_i et S_j. Nous avons choisi un calcul à base de différence mais le calcul sous forme d'un rapport est aussi envisageable. De plus, la différence entre les valeurs d'expression observées pour un gène g en situation S_j par rapport à S_i est ici normalisée par la division par le minimum des valeurs d'expression observées pour ce gène dans l'ensemble des situations considérées.
Soit G l'ensemble des gènes du jeu de données, et n le nombre de situations S_i étudiées. Pour tout gène g appartenant à G, nous notons ν_{s_i} la valeur d'expression de g dans la situation S_i. Nous définissons alors la variation d'expression de g pour la paire de situation (S_i, S_j) par l'équation suivante :

$$\xi_{i,j}(g) = \frac{\nu_{s_i} - \nu_{s_j}}{\mathbf{min_{k=1..n}}\ \nu_{sk}}. \quad (1)$$

La modélisation floue conduit à définir une valeur seuil σ pour les valeurs prises par $\xi_{i,j}(g)$, et, pour chaque paire de situations (S_i,S_j), trois fonctions de degré d'appartenance à des ensembles

flous permettent de définir le comportement des gènes en fonction des valeurs prises par $\xi_{i,j}$. Ces trois fonctions sont décrites ci-dessous :

$Over_{i,j}$: G $\longrightarrow$ [0,1]

$$Over_{i,j}(g) = \begin{cases} 1 & si \quad \xi_{i,j}(g) \geq \sigma \\ \frac{1}{\sigma} \times \xi_{i,j}(g) & si \quad 0 < \xi_{i,j}(g) < \sigma \\ 0 & sinon \end{cases}$$

$Under_{i,j}$: G $\longrightarrow$ [0,1]

$$Under_{i,j}(g) = \begin{cases} 1 & si \quad \xi_{i,j}(g) \leq -\sigma \\ \frac{-1}{\sigma} \times \xi_{i,j}(g) & si \quad -\sigma < \xi_{i,j}(g) < 0 \\ 0 & sinon \end{cases}$$

$Iso_{i,j}$: $G \longrightarrow$ [0,1]

$$Iso_{i,j}(g) = \begin{cases} 0 & si \quad \xi_{i,j}(g) < -\sigma \quad ou \quad \xi_{i,j}(g) > \sigma \\ \frac{1}{\sigma} \times \xi_{i,j}(g) + 1 & si \quad -\sigma < \xi_{i,j}(g) < 0 \\ \frac{-1}{\sigma} \times \xi_{i,j}(g) + 1 & si \quad 0 < \xi_{i,j}(g) < \sigma \end{cases}$$

La figure (26) présente les fonctions $Over_{i,j}$, $Under_{i,j}$, $Iso_{i,j}$. La fonction $Over_{i,j}(g)$ permet d'affecter un gène g à l'ensemble d'expression différentielle $\text{Over}_{i,j}$, avec un degré d'appartenance équivalent à 1, si la différence d'expression de ce gène $\xi_{i,j}(g)$ par rapport aux deux situations S_i et S_j est supérieur à σ. Si la valeur $\xi_{i,j}(g)$ appartient à l'intervalle]0,σ[alors le gène sera affecté à la fois à l'ensemble $\text{Over}_{i,j}$ avec un degré d'appartenance de $\frac{1}{\sigma} \times \xi_{i,j}(g)$ et à l'ensemble $\text{Iso}_{i,j}$ avec un degré d'appartenance de $\frac{-1}{\sigma} \times \xi_{i,j}(g) + 1$. Dans ce cas, il paraît clairement que si la valeur de différence d'expression est dans l'intervalle [0,σ] alors le gène aura une double appartenance aux deux ensemble $\text{Iso}_{i,j}$ et $\text{Over}_{i,j}$, exprimant ainsi l'intérêt d'une modélisation floue. La même analyse se fait pour les autres valeurs prises par $\xi_{i,j}(g)$ par rapport aux autres ensembles d'expression différentielle. Le processus de "*defuzzification*" conduit ensuite à l'affectation définitive d'un gène g à un ou plusieurs ensembles d'expression différentielle en fonction d'une valeur seuil μ selon le principe suivant : si la valeur prise par la fonction $Over_{i,j}(g)$, $Under_{i,j}(g)$ ou $Iso_{i,j}(g)$ est supérieur ou égale à μ, alors le gène est affecté à l'ensemble correspondant.
Prenons un exemple (Figure 26), pour un gène g, on a $Over_{i,j}(g) = 0,25$, $Under_{i,j}(g) = 0$, $Iso_{i,j}(g) = 0,75$. On voit que pour $\mu = 0,5$, le gène g sera affecté à l'ensemble $Iso_{i,j}$ uniquement, alors que pour $\mu = 0,25$ ce gène sera affecté à la fois à l'ensemble $Iso_{i,j}$ avec un degré d'appartenance de 0.75 et $Over_{i,j}$ avec un degré d'appartenance complémentaire de 0,25. On remarque aussi qu'avec ce type de modèle, la bi-appartenance d'un gène à deux ensembles d'expression différentielle n'est possible que pour $\mu \leq 0,5$. Par extension, un profil d'expression différentielle étant défini comme une combinaison d'ensembles d'expression différentielle, la modélisation floue introduite ici permet aussi que certains gènes puissent appartenir à plus d'un profil.

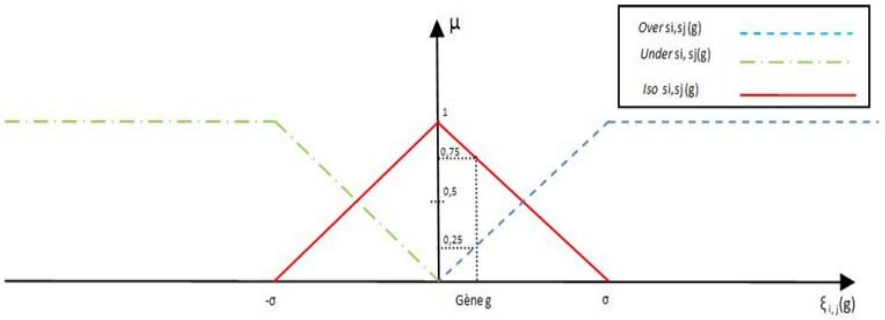


FIGURE 26 – Représentation des trois fonctions de degré d'appartenance d'un gène g aux ensembles d'expression différentielle $Over_{i,j}$, $Under_{i,j}$, $Iso_{i,j}$ utilisés pour définir les profils.

2 Exemple applicatif sur une liste de gènes du cancer colorectal différentiellement exprimés

2.1 Calcul des profils d'expression différentielle

Nous rappelons que nous avons en entrée une liste de 222 gènes avec leurs valeurs d'expression dans 3 situations : tissu colo-rectal sain (S_1), tumeur colo-rectale (S_2), lignée cellulaire cancéreuse en culture (S_3).
La combinaison des 3 ensembles d'expression différentielle possibles pour chacune des 3 paires de situations conduit à un nombre total de 27 profils possibles, comme illustré dans la Table 16.
La mise en œuvre de la méthode d'analyse selon le modèle flou décrit ci-dessus nécessite de choisir les valeurs des deux paramètres σ et μ.
Le choix des deux paramètres σ et μ, dépend du jeu de données considéré. Une analyse exploratoire a donc été réalisée avec le jeu de données étudié en faisant varier les deux valeurs (μ={0,1, 0,2, 0,3, 0,4, 0,5} ; σ={0,1, 0,2, 0,3, 0,4, 0,5, 0,6, 0,7, 0,8, 0,9, 1}). Un programme en C++ a été implémenté et les profils d'expression différentielle ont été calculés pour chaque combinaison de μ et σ et le nombre de gènes présents dans chaque profil a été déterminé [BDST+09]. La figure 27 donne un exemple de distribution des gènes dans les 27 profils pour $\mu = 0,3$ et toutes les valeurs de σ. Les critères utilisés par les biologistes lors de la sélection du jeu de données permettent de formuler des contraintes conduisant à éliminer certaines valeurs de σ et μ. En effet, les 222 gènes retenus pour cette étude présentent une dérégulation significative dans la tumeur colo-rectale par rapport au tissu normal ($Over_{2,1}$ ou $Under_{2,1}$). Les profils contenant l'ensemble d'expression différentielle $Iso_{2,1}$ doivent donc être dépourvus de gènes. Ce critère nous amène à éliminer les valeurs de $\sigma \geq 0,6$, ceci pour toutes les valeurs de μ, car ces combinaisons génèrent des profils contenant l'ensemble $Iso_{2,1}$.
Avec l'avis de l'expert, nous avons décidé de continuer le travail d'analyse en gardant la combinaison (μ,σ)=(0,3, 0,4).
La Table 17 illustre, sur deux exemples de gènes, la procédure permettant de déterminer les profils d'expression des gènes en calculant les valeurs de leurs degrés d'appartenance aux 3 ensembles d'expression différentielle pour chaque paire de situations et pour les valeurs de (μ,σ)=(0,3, 0,4). On constate que le gène ($g2$) est classifié dans les profils Profil-2 et Profil-20, du fait des valeurs $Over_{3,1}(g2) = 0,38$ et $Iso_{3,1}(g2) = 0,62$. Cet exemple illustre l'intérêt de la modélisation floue pour des valeurs limites où il est difficile de trancher l'appartenance d'un gène à tel ou tel en-

	Ensembles d'Expression Différentielle								
	$Over_{3,1}$	$Under_{3,1}$	$Iso_{3,1}$	$Over_{2,1}$	$Under_{2,1}$	$Iso_{2,1}$	$Over_{3,2}$	$Under_{3,2}$	$Iso_{3,2}$
Profil1	×			×			×		
Profil2	×			×				×	
Profil3	×			×					×
Profil4	×				×		×		
Profil5	×				×			×	
Profil6	×				×				×
Profil7	×					×	×		
Profil8	×					×		×	
Profil9	×					×			×
Profil10		×		×			×		
Profil11		×		×				×	
Profil12		×		×					×
Profil13		×			×		×		
Profil14		×			×			×	
Profil15		×			×				×
Profil16		×				×	×		
Profil17		×				×		×	
Profil18		×				×			×
Profil19			×	×			×		
Profil20			×	×				×	
Profil21			×	×					×
Profil22			×		×		×		
Profil23			×		×			×	
Profil24			×		×				×
Profil25			×			×	×		
Profil26			×			×		×	
Profil27			×			×			×

TABLE 16 – Liste des 27 profils possibles obtenus avec la combinaison des 3 ensembles d'expression différentielle possibles pour chacune des 3 paires de situations.

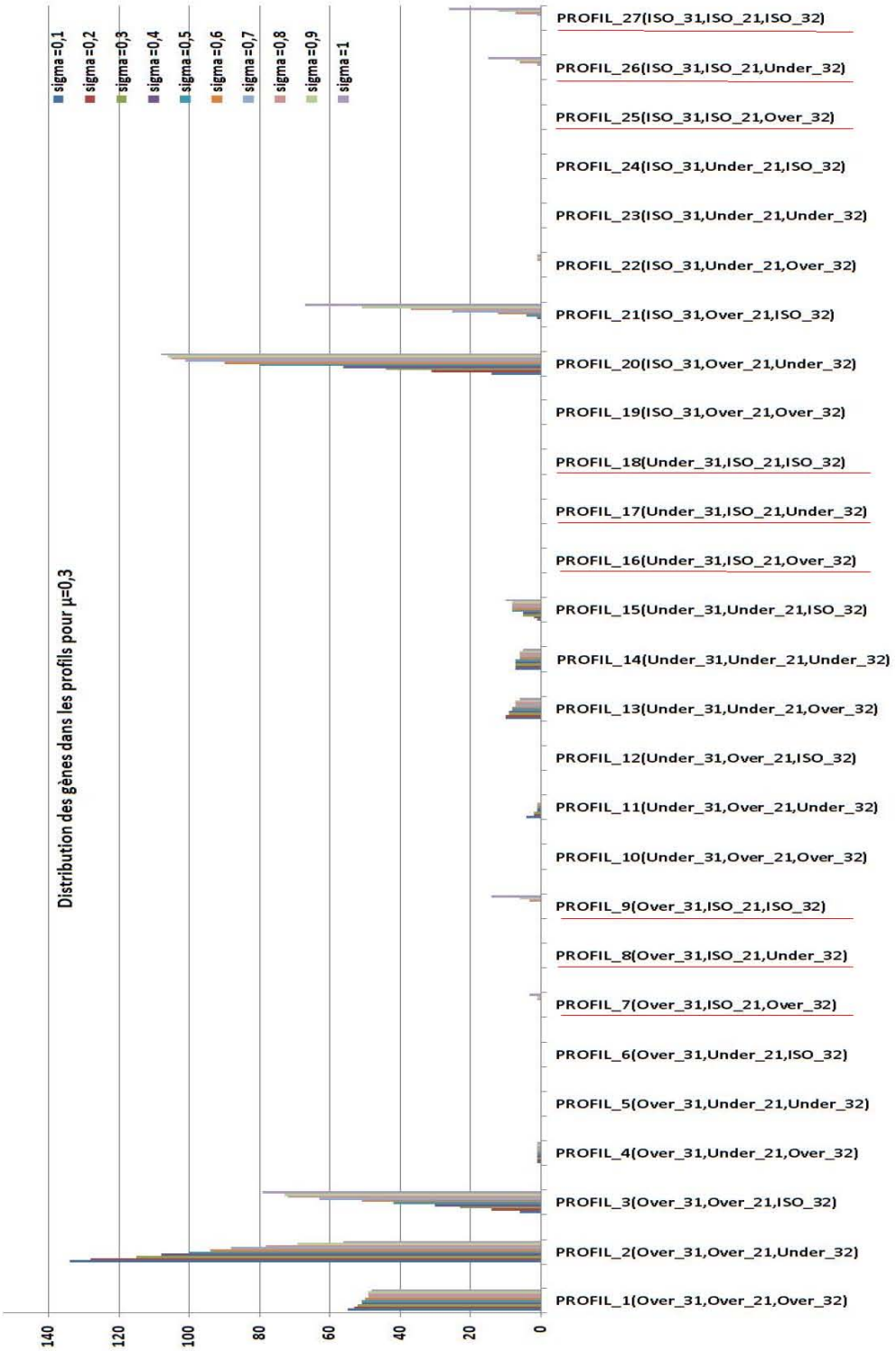


FIGURE 27 – Exemple de distribution des 222 gènes du jeu de données relatif au cancer colorectal dans les profils pour μ=0,3, et pour des valeurs de σ entre 0,1 et 1. Les profils contenant l'ensemble $Iso_{2,1}$ sont soulignés en rouge pour aider à repérer les profils dans lesquels on ne devrait pas avoir de gènes, suite à la sélection initiale des 222 gènes.

semble d'expression différentielle.

	g1	*g2*
ν_{s_1}	33.6	22.1
ν_{s_2}	827.9	70.7
ν_{s_3}	735.8	25.4
$\xi_{3,1}(g)$	20.90	0.15
$\xi_{2,1}(g)$	23.64	2.20
$\xi_{3,2}(g)$	-2.74	-2.05
calcul des degrés d'appartenance pour $\sigma = 0,4$		
$Over_{3,1}(g)$	1	0.38
$Under_{3,1}(g)$	0	0
$Iso_{3,1}(g)$	0	0.62
$Over_{2,1}(g)$	1	1
$Under_{2,1}(g)$	0	0
$Iso_{2,1}(g)$	0	0
$Over_{3,2}(g)$	0	0
$Under_{3,2}(g)$	1	0
$Iso_{3,2}(g)$	0	1
Profil(s) pour pour μ=0,3	Profil1($Over_{3,1}$,$Over_{2,1}$,$Over_{3,2}$)	Profil2($Over_{3,1}$,$Over_{2,1}$,$Under_{3,2}$) Profil20($Iso_{3,1}$,$Over_{2,1}$,$Under_{3,2}$)

TABLE 17 – Exemples de calculs pour la détermination des profils de deux gènes $g1$ et $g2$. La définition des profils à partir des différentes combinaisons d'ensembles d'expression différentielle est donnée dans la table 16.

Le programme d'affectation des gènes aux profils d'expression différentielle a été exécuté sur l'ensemble du jeu de données pour (μ,σ)=(0,3, 0,4). Il fournit les résultats présentés dans la Table 18. Les gènes se répartissent dans 10 des 27 profils possibles selon une distribution correspondant à la diagonale de la matrice présentée sur la Tableau 18. Un recouvrement est observé entre certains profils en terme de nombre de gènes partagés.

	Profil1	Profil2	Profil3	Profil4	Profil11	Profil13	Profil14	Profil15	Profil20	Profil21
Profil1	51		2							
Profil2		108	17						24	1
Profil3			30						1	1
Profil4				1						
Profil11					1					
Profil13						9		1		
Profil14							7	1		
Profil15								5		
Profil20									56	
Profil21										1

TABLE 18 – Matrice de distribution des gènes dans les profils. Notons que si une cellule est vide alors les deux profils correspondants ne partagent aucun gène. La diagonale représente le nombre de gènes de chaque profil.

J'ai ensuite décidé de réaliser une analyse fonctionnelle exploratoire sur les gènes des profils d'expression différentielle du jeu de données relatif au cancer colo-rectal, en utilisant les outils de classification DAVID et la mesure de similarité sémantique *IntelliGO*. Le but de cette analyse est d'associer aux différents profils d'expression, des fonctions biologiques significatives qui sont spécifiques de ces profils. Les résultats de ces analyses sont présentés dans les sous sections suivantes.

2.2 Exploration des annotations fonctionnelle des profils d'expression différentielle avec l'outil DAVID

Appliqué sur la totalité des gènes du jeu de données, et sur les gènes des profils définis précédemment, l'outil de classification des annotations fonctionnelle DAVID donne les résultats consignés dans la Table 19. Dans cette analyse, les annotations fonctionnelles de tous les gènes d'un profil sont clusterisées par l'outil DAVID et permettent d'isoler des groupes de fonctions biologiques dans chaque profil, groupes ou clusters caractérisés par un score d'enrichissement dont les éléments (annotations GO ou autres) sont associés à une P-value.
L'analyse a fourni pour les 222 gènes du jeu de données complet une liste ordonnée de 51 clusters

	Nombre de clusters d'annotation fonctionnelle	*scores min* *scores min*	*score max* *score max*	*score moy.* *score moy.*	*M.V.S.* *M.V.S.*
Jeu de données complet (222)	51	0.02	1.66	0.56	9.50E-04
Profil1 (51)	17	0.07	2.51	0.78	5.3E-04
Profil2 (108)	31	0.09	1.48	0.5	3.0E-03
Profil3 (30)	7	0.11	0.6	0.36	5.7E-2
Profil13 (9)	3	0.06	0.67	0.3	6E-02
Profil14 (7)	4	0.06	0.86	0.48	3.2E-02
Profil15 (5)	1	0.19	0.19	0.19	2.3E-01
Profil20 (56)	18	0.03	1.51	0.48	1E-03

TABLE 19 – Résultats du clustering d'annotation fonctionnelle (DAVID) sur tous les gènes du jeu de données du cancer (première ligne), et sur les gènes appartenant aux profils. Chaque ensemble est suivi du nombre de gènes qu'il contient. *M.V.S.* :meilleure valeur de significativité correspondant à la plus petite P-Value. Les scores (Min, Max, Moyen) représentent des agrégations (en Log) des valeurs p-value des annotations fonctionnelles dans les ensembles de gènes considérés.

d'annotations fonctionnelles dont les premiers ont des scores d'enrichissement tout à fait acceptables (supérieur à 1) et contiennent des termes d'annotation dont l'occurrence est significative (P-value $\leq$ 10E-04). Pour les profils pris séparément le nombre de clusters obtenus augmente en fonction du nombre de gènes présents dans le profil avec un maximum de 31 clusters pour le profil Profil2 qui contient 108 gènes et un minimum de 1 cluster pour le profil Profil15 qui ne contient que 5 gènes. D'un point de vue quantitatif on s'aperçoit que les scores des clusters peuvent atteindre, pour les profils, des valeurs sensiblement plus élevées que celles obtenues avec le jeu de données complet.
Afin d'estimer la contribution de la répartition des gènes dans des profils d'expression différentielle à cette analyse fonctionnelle, il a été demandé à l'expert biologiste de regarder en détail les contenus des clusters d'annotation fonctionnelle et d'établir une correspondance entre les clusters obtenus pour chaque profil et ceux obtenus pour l'ensemble du jeu de données.
Les résultats sont présentés dans la Table 20. Les numéros sont ceux des clusters d'annotation fonctionnelle obtenus par l'analyse DAVID pour le jeu de données complet (1ère ligne) et pour chacun des profils d'expression différentielle étudiés (lignes suivantes). Le recouvrement complet ou partiel des clusters obtenus pour les profils avec ceux obtenus pour le jeu de données complet est indiqué par le numéro du cluster dans la colonne correspondante. En jaune sont indiqués les clusters du jeu de données complet pour lesquels aucun recouvrement n'a été trouvé avec les clusters des profils, en vert les clusters du jeu de données complet qui ont été retrouvés parmi les clusters d'un seul profil. La colonne "*Nouveau*" indique que les clusters observés n'ont pas de recouvrement avec les clusters du jeu de données complet.

Plusieurs conclusions peuvent être tirées de cette analyse. Tout d'abord il apparaît que cer-

	Numéro du cluster d'annotation fonctionnelle (partie 1/4)														
Jeu de données complet	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Profil1		3	1							5/12		3	12		
Profil2	1		18	1		7/19	3	6		14	9				8
Profil3															
Profil13															
Profil14								2							
Profil15															
Profil20	1/3			1				2/5	3	1	8			5	

	Numéro du cluster d'annotation fonctionnelle (partie 2/4)													
Jeu de données complet	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Profil1			10	15					4,8	7	6			1/5/16
Profil2	4	2		11		10		13	27	31			24	20
Profil3				5		6				1/2		7		
Profil13				1										
Profil14				1/3		1								
Profil15														
Profil20		12		2/14/15		7				16	6	15		

tains clusters d'annotation fonctionnelle observés lors de l'analyse du jeu de données complet sont spécifiques à l'un des profils déterminés. C'est le cas de 20 / 51 soit 40 %, environ des clusters du jeu de données complet (colorés en vert). Ainsi il devient possible d'associer un comportement transcriptionnel particulier, pour un contexte de situations donné, à ces descripteurs de fonction biologique.
A contrario, on remarque que certains clusters du jeu de données complet se retrouvent dans des clusters associés à plus d'un profil d'expression (21 sur 51 soit 40%). Les fonctions biologiques décrites dans ces clusters ne sont donc pas associées à un comportement transcriptionnel particulier des gènes correspondants. Cette observation peut cependant résulter de la modélisation floue des profils et une interprétation transcriptionnelle peut être établie dans le cas où deux profils partageant des gènes sont concernés par le même cluster du jeu de données complet.
De façon intéressante certains clusters du jeu de données complet ne présentent aucun recouvrement avec les clusters obtenus dans les profils (colorés en jaune). C'est le cas de 10 clusters sur 51 soit environ 20 %. Une interprétation possible est que les fonctions biologiques décrites dans ces clusters sont associées à des gènes distribués dans des profils différents de telle sorte qu'il n'est pas possible de voir émerger ces descripteurs en prenant les profils séparément. Cette observation mérite d'être explorée plus avant car elle met à l'épreuve l'hypothèse classique selon laquelle les gènes partageant une même fonction biologique partagent aussi le même profil d'expression.
Enfin il est également intéressant de constater que l'analyse fonctionnelle par profil fait émerger des clusters d'annotation fonctionnelle qui ne sont pas retrouvés parmi les clusters du jeu de données complet (colonne "Nouveau"). Ces clusters regroupent des termes d'annotation qui devaient être trop minoritaires dans le jeu de données complet.

L'analyse fonctionnelle avec l'outil DAVID nous a permis d'explorer les PED et de produire des résultats facilement interprétables par l'expert biologiste. Afin de valider les résultats rapportés ici, il sera intéressant de tester d'autres programmes d'annotation fonctionnelle. Afin d'étendre notre approche, nous voudrions mieux formaliser l'interprétation des profils d'expression à partir des résultats de l'annotation fonctionnelle en tenant compte des relations entre les éléments d'annotation fonctionnelle. En effet, comme nous l'avons introduit dans le premier chapitre, l'outil d'analyse fonctionnelle DAVID se base sur la statistique Kappa, et ne prend pas en considération les relations sémantiques entre les termes d'annotations fonctionnelles. Or nous avons constaté grâce à l'évaluation de la classification fonctionnelle de gènes au chapitre 3 que la classification obtenue avec notre mesures de similarité sémantique *IntelliGO* est robuste par

	Numéro du cluster d'annotation fonctionnelle (partie 3/4)												
Jeu de données complet	30	31	32	33	34	35	36	37	38	39	40	41	42
Profil1					15	11			7	13	17		
Profil2		12	5			21/35		26			2		
Profil3											3		
Profil13													
Profil14													
Profil15			4										
Profil20					13	10					18		

	Numéro du cluster d'annotation fonctionnelle (partie 4/4)									
Jeu de données complet	43	44	45	46	47	48	49	50	51	Nouveau
Profil1	2/4			14						
Profil2	23/28		15	17			29	22	30	16
Profil3	3			3						4
Profil13	2/3									
Profil14	3/4									
Profil15	1									
Profil20	11/16/17			12		7/9				

TABLE 20 – Distribution des clusters d'annotation fonctionnelle à travers les profils d'expression différentielle (suite).

rapport à la classification fonctionnelle DAVID. De plus, notre méthode de classification exploite, par le biais de la définition de la mesure de similarité sémantique, les relations hiérarchiques qui existent entre les annotations. Nous avons décidé d'utiliser le clustering fonctionnel basé sur la mesure *IntelliGO* afin de guider le processus d'analyse fonctionnelle sur les données d'expression, et de démontrer l'intérêt de la méthode de l'analyse de recouvrement avec les profils d'expression différentielle comme ensembles de référence (cf chapitre 3).

2.3 Analyse de recouvrement entre profils d'expression différentielle et clusters fonctionnels IntelliGO

La classification fonctionnelle basée sur *IntelliGO* est utilisée ici sur les données d'expression du cancer colo-rectal. Un ensemble de 128 gènes sont gardés à partir de la liste initiale des 222 gènes, du fait de leurs annotations fonctionnelles sur l'aspect Biological Process de l'ontologie GO. Le principe général ici est de confronter les clusters fonctionnels obtenus avec *IntelliGO* en utilisant la liste de 128 gènes, aux profils d'expression différentielle flous obtenus avec la même liste de gènes. Intuitivement, l'analyse de recouvrement devrait découvrir des relations implicites entre des comportements transcriptionnels et des fonctions biologiques.
En reprenant l'algorithme (2), nous considérons ici les profils d'expression différentielle flous (fuzzy DEPs) comme la collection d'ensembles de référence Σ. Plus précisément, nous gardons huit fuzzy DEPs représentatifs des 128 gènes du cancer colo-rectal, tels que $\Sigma = \{Profil1, Profil2, Profil3, Profil13, Profil14, Profil20, Profil22\}$. Tout d'abord, une matrice de similarité *IntelliGO* est calculée entre toutes les paires de gènes de la liste des 128 gènes considérés ici. Comme étape préliminaire, nous avons utilisé cette matrice de similarité pour réaliser un clustering hiérarchique avec une visualisation par heatmap afin d'estimer la distribution des gènes dans les clusters fonctionnels. Les résultats sont affichés dans la figure (28). Malgré la variation de l'intensité des couleurs dans les différentes régions de la heatmap, plusieurs clusters peuvent être distingués dans sa diagonale. Le cluster très homogène situé dans la partie supérieure gauche de la heatmap et présentant très peu de similarité croisée avec les autres gènes du jeu de données a été étudié du point de vue de son contenu en gènes. Les résultats concernant les 11 gènes de ce cluster sont résumés dans l'Annexe B. La forte similarité fonctionnelle existant entre ces gènes s'explique par des annotations BP récurrentes concernant des processus de transport. Les

processus de transport sont importants dans la physiologie de l'appareil digestif et certains gènes du cluster sont déjà connus pour être dérégulés dans le cancer colorectal. C'est le cas en particulier du gène AQP8 de l'aquaporine 8 dont l'expression n'est plus détectable dans les tumeurs colorectales [FSRL01]. De fait le gène AQP8 est retrouvé dans le PED Profil_14 qui correspond aux gènes sous-exprimés dans la situation tumeur par rapport à la situation saine. Un autre gène (ATP11A) est aussi présent dans le cluster fonctionnel analysé ici, avec une annotation de transport de phospholipides, et a été décrit récemment comme nouveau marqueur prédictif de métastases métachroniques du cancer colorectal [MIM+10]. Associé dans notre étude au PED Profil_1, il apparaît effectivement comme sur-exprimé dans la situation tumeur par rapport à la situation saine. Ces deux gènes peuvent être considérés comme des témoins positifs, validant le bien-fondé de notre approche de recouvrement entre PEDs et clusters fonctionnels IntelliGO. Le rôle possible des neuf autres gènes du cluster doit encore être exploré par les experts biologistes qui sont à l'origine de l'expérience de transcriptomique analysée par l'IGBMC.

Afin d'étendre notre analyse, nous avons appliqué un clustering flou en utilisant *IntelliGO*, en

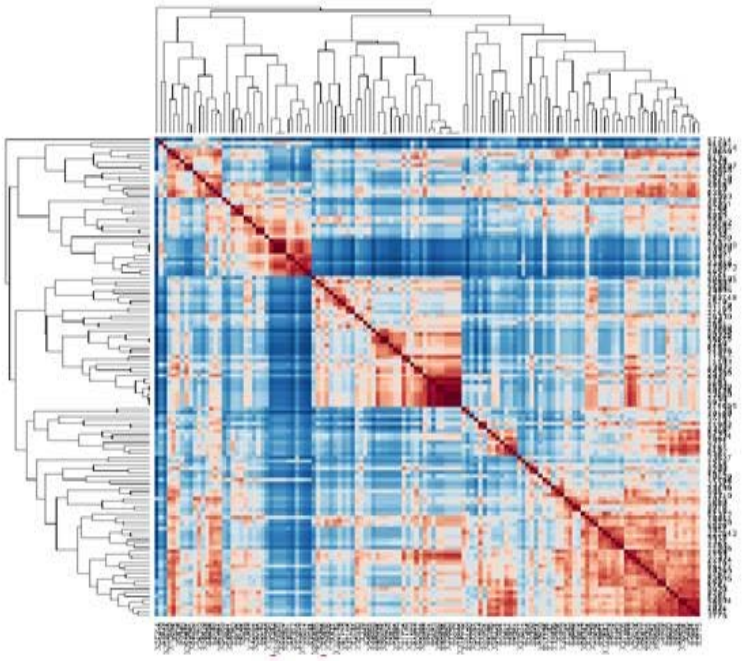


FIGURE 28 – Une Heatmap générée avec les paires de similarité IntelliGO sur les 128 gènes du cancer colo-rectal.

partant de nos hypothèses évoquées et argumentées précédemment. Pour déterminer le nombre optimal de clusters fonctionnels sur ces 128 gènes, nous avons utilisé l'algorithme (2). Dans ce cas là, nous avons varié le nombre k de clusters à générer dans l'intervalle [2,14], sachant que le nombre d'ensembles de référence est huit fuzzy DEPs. La Table 21 illustre les résultats obtenus avec cet algorithme. Le nombre optimal de clusters k est de 3 avec un F-score global de $0,40$.

Par la suite, l'analyse de recouvrement avec l'algorithme (3) nous fournit une liste des paires les mieux appariées entre les 3 clusters fonctionnels et les 8 PEDs. Dans ce cas, une paire (R,C) représente les gènes qui sont à la fois fonctionnellement similaires et présents dans un même profil d'expression $C \cap R$. La liste de ces paires et le nombre de gènes qu'elles contiennent sont

K clusters fonctionnels obtenus avec IntelliGO	2	3	4	5	6	7	8	9	10	11	12	13	14
Global F-Score	0.39	0.4	0.37	0.32	0.26	0.27	0.29	0.29	0.29	0.28	0.26	0.25	0.16

TABLE 21 – Variation du F-Score global pour les différentes valeurs du nombre k de clusters à générer dans l'intervalle [2,14]. Le clustering flou est utilisé sur les 128 gènes du cancers colorectal annotés avec les annotations BP de l'ontologie GO. Les 8 fuzzy DEPs extraits à partir de ces 128 gènes sont considérés ici comme ensembles de référence pour le calcul des valeurs du F-score.

présentés dans la Table 22.
Par exemple, les méthodes d'enrichissement statistique permettent d'extraire à partir des anno-

	Fuzzy Differential Expression Profiles							
	Profil1 (34)	Profil2 (51)	Profil3 (32)	Profil13 (6)	Profil14 (5)	Profil15 (4)	Profil20 (31)	Profil22 (1)
Cluster1 (45)			14	3				1
Cluster2 (64)	15	28				2	17	
Cluster3 (19)					3			

TABLE 22 – L'analyse de recouvrement entre trois clusters fonctionnels obtenus avec le clustering flou *IntelliGO*, et les huit fuzzy DEPs. Entre parenthèses est affiché le nombre de gènes que contient chaque ensemble. Pour chaque paire d'ensembles appariés, le nombre de gènes présents dans l'intersection $C \cap R$ est affiché dans la cellule.

tations GO qui annotent tous les gènes d'un cluster donné, les termes qui sont statistiquement signifiants ayant des P-Value $< 10^{-4}$ ou 10^{-5}. Rappelons que les valeurs P-value sont calculées à partir d'une liste de gènes d'intérêt par rapport à une liste globale de référence (background). Dans notre cas, nous avons utilisé le test hypergéométrique [ENS+09], appliqué sur les termes GO qui annotent les gènes présents dans $C \cap R$ c'est à dire un cluster et un fuzzy DEP. Les gènes du corpus NCBI [NCB], sont utilisés ici pour former la liste globale de gènes de référence. Les résultats sont présentés dans la Table 23 pour chaque paire (C,R). Nous pouvons remarquer que des termes GO très spécifiques de l'aspect BP, sont affectés aux gènes de l'intersection $C \cap R$ d'un cluster fonctionnel *IntelliGO* et d'un profil d'expression fuzzy DEP. Dans le cas du Cluster_2 ∩ Profil2 et Cluster_2 ∩ Profil20, un terme GO assez général (*cell differentiation*) est obtenu dans la première position, mais différents termes GO apparaissent en seconde et troisième position. Ceci correspond aux processus biologiques qui sont mélangés ensemble dans le Cluster 2 mais ont été par la suite associés à deux PEDs différent (Profil2 et Profil20). Cet exemple illustre bien que notre analyse de recouvrement est capable d'extraire des sous ensembles consistants de gènes, caractérisés par des fonctions biologiques spécifiques et un comportement transcriptomique particulier [BSTP+ce]. Des résultats plus détaillés sont présentés dans l'Annexe C.

3 Discussion

Notre proposition de modélisation a priori de profils d'expression différentielle sur la base d'une modélisation floue, offre une alternative intéressante aux méthodes traditionnelles lorsque la nature des relations entre situations l'exige. Elle permet en particulier de regrouper ensemble des gènes dont l'expression varie de la même façon entre deux situations indépendamment de la valeur de leur niveau d'expression dans chaque situation. Dans une analyse classique, ces gènes se seraient retrouvés dans des profils différents. L'analyse fonctionnelle réalisée ici montre de

deux façons (DAVID et clustering flou avec *IntelliGO*) que l'introduction des profils d'expression différentielle permet de mettre en relation des fonctions biologiques et des comportements transcriptionnels particuliers.

Cluster_1 ∩ Profil3		Cluster_2 ∩ Profil1		Cluster_2 ∩ Profil2		Cluster_2 ∩ Profil20		Cluster_3 ∩ Profil14	
terme GO	P-Value	terme GO	P-Value	terme GO	P-Value	terme GO	P-Value	terme GO	P-Value
regulation of transcription, DNA-dependent	9.95E-04	chromosome organization	9.55E-05	cell differentiation	7.35E-05	cell differentiation	5.97E-05	Water transport	2.08E-05
NADH oxidation	2.98E-04	strand break repair via homologous recombination	1.1012E-04	vascular endothelial growth factor receptor signaling pathway	1.06E-04	multicellular organismal development	9.58E-05		
		response to estrogen stimulus	9.6584E-04	angiogenesis	5.83E-04	insulin secretion	1.42E-04		

TABLE 23 – L'analyse d'enrichissement des termes GO pour les paires (R,C) les mieux appariées. Nous n'affichons ici que les termes GO les plus spécifiques avec une P-Value inférieure à 10^{-3} qui caractérisent les gènes présents dans l'intersection $C \cap R$.

Chapitre 5

Vers une extension de la mesure de similarité sémantique à d'autres ontologies

Sommaire

1 Généralisation de la mesure IntelliGO à des graphes dirigés acycliques enracinés (rDAG)

1.1 Motivation

Le problème de la généralisation de la mesure de similarité sémantique IntelliGO m'a été posé à propos du travail d'un autre doctorant dans l'équipe Orpailleur (Emmanuel Bresso). Ce doctorant s'intéressait à l'étude des effets secondaires des médicaments. Il voulait mettre en oeuvre des méthodes des fouille de données sur un jeu de données représentant des molécules (objets), associées à leurs effets secondaires (attributs). Or les termes décrivant les effets secondaires étaient d'une très grande diversité, ce qui rendait peu efficaces les algorithmes de fouille de données symboliques utilisés (motifs fréquents, recherche de sous-groupes). Remarquant que ces termes provenaient d'un vocabulaire (MedDRA, voir plus loin), structuré en DAG enraciné comme l'ontologie GO, Emmanuel Bresso m'a demandé s'il ne serait pas possible d'étendre la

mesure de similarité terme-terme *IntelliGO* à ce vocabulaire afin de réaliser un clustering des termes décrivant les effets secondaires.
L'idée générale dans cette partie de travail est donc d'exploiter les relations sémantiques existant entre les termes décrivant les effets secondaires, pour réduire les dimensions du jeu de données en réalisant un clustering sémantique des attributs. Ceci m'a conduit à modifier légèrement la définition de la mesure de similarité sémantique *IntelliGO* dans sa partie terme-terme, pour la rendre plus générale quel que soit le vocabulaire où on l'applique.

1.2 Introduction de la similarité sémantique généralisée SimDAG

La mesure de similarité sémantique *IntelliGO* définie au chapitre 2, dans l'équation (18) exprime la proximité sémantique de deux termes de l'ontologie GO. Cette formule peut être étendue à toute paire de concepts dans une ontologie représentée sous forme de rDAG. De plus une légère modification peut être apportée à la formule en considérant la situation suivante. Dans la Figure 29, nous distinguons dans cet exemple de rDAG, deux chemins qui relient les

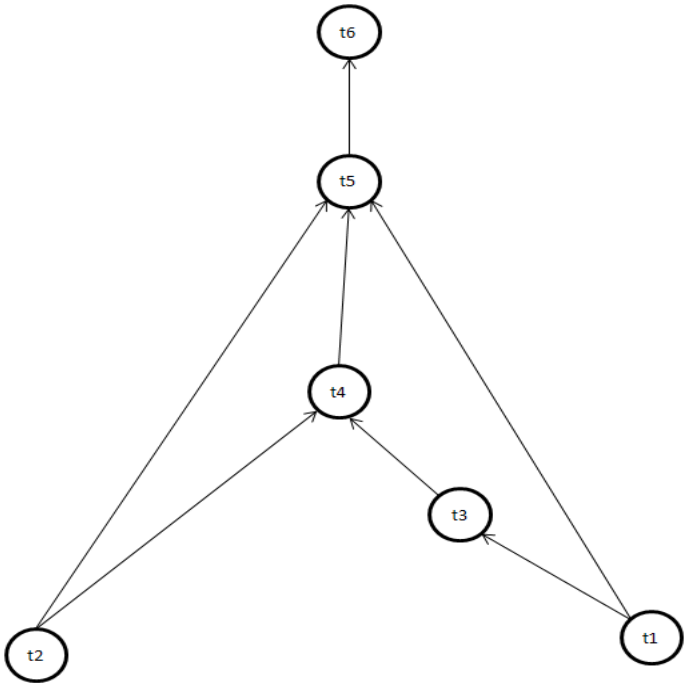


FIGURE 29 – Illustration du calcul du plus court chemin local qui passe par un LCA, et du plus court chemin global qui ne passe pas obligatoirement par un LCA.

deux termes t_1 et t_2 : le premier chemin est ϕ_1= $\{t_1,t_3,t_4,t_2\}$ qui passe par un LCA des deux termes qui est le terme t_4, tandis que le deuxième chemin est ϕ_2= $\{t_1,t_5,t_2\}$ qui passe par un ancêtre commun t_5 qui n'est pas un LCA. Dans la définition précédente de notre similarité sémantique de l'équation (18), le chemin considéré pour calculer la valeur $MinSPL(t_1, t_2)$ est ϕ_1 car il s'agit du SPL minimum passant par un LCA, puisque $LCA(t_1, t_2) \in \phi_1$. Par conséquent la valeur $MinSPL(t_1, t_2)$ est égale à $Length(\phi_1)$= 3. Or si on considère le deuxième chemin, nous avons $Length(\phi_2)$= 2. De fait, on voit que l'on peut trouver dans un rDAG un chemin reliant deux nœuds, qui est plus court que celui qui passe par un des LCA. Ignorer cette possibilité revient à considérer au dénominateur de la formule (18) une valeur $MinSPL(t_i, t_j)$ qui n'est pas le minimum des chemins entre t_i et t_j. L'existence d'un SPL plus court que celui passant par un

LCA renforce la similarité des deux termes comparés, ce qui suggère de maximiser la mesure de similarité en remplaçant $MinSPL(t_i, t_j)$ par un SPL tout simple. On obtient alors la formule :

$$SimDAG(t_i, t_j) = \frac{2 * Depth(LCA((t_i, t_j)))}{SPL(t_i, t_j) + 2 * Depth(LCA((t_i, t_j)))}, \quad (1)$$

et la distance associée est :

$$DistDAG(t_i, t_j) = \frac{SPL(t_i, t_j)}{SPL(t_i, t_j) + 2 * Depth(LCA((t_i, t_j)))}. \quad (2)$$

1.3 Application à deux ontologies de pharmacovigilance COSTART et MedDRA

Afin d'appliquer la nouvelle similarité sémantique $SimDAG$, nous avons utilisé deux ontologies dont la structure est représentée sous forme de rDAG. La première ontologie est COSTART : The Coding Symbols for a Thesaurus of Adverse Reaction Terms, qui est un vocabulaire faisant partie du méta-thésaurus UMLS. Les termes d'annotation de l'ontologie COSTART ont été proposés par l'hôpital du Massachusetts dans le but de la gestion des aliments et l'annotation de leurs effets secondaires [Cam86, Hat79]. La deuxième ontologie est MedDRA : Medical Dictionary for Drug Regulatory Activities, qui est une terminologie de pharmacovigilance pour coder les actions et les effets secondaires des molécules de médicaments [Mer08] [13]. L'intérêt de telles ontologies est dans leur application dans les processus de soin de patients en améliorant la disponibilité de l'information médicale en termes d'accessibilité, de rapidité de la récupération, de lisibilité, et d'organisation. Un extrait du graphe de l'ontologie MedDRA est donné dans la Figure 30. Il est

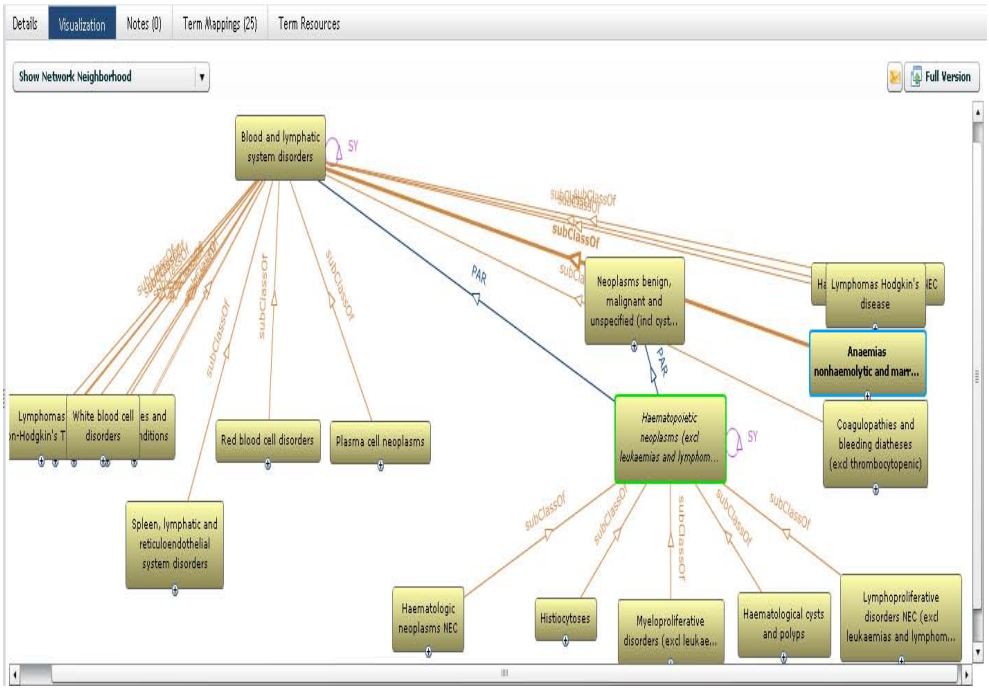


FIGURE 30 – Extrait d'un sous-graphe de l'ontologie MedDRA.

13. http ://bioportal.bioontology.org/ontologies/42280 ?p=terms

à noter que tout le vocabulaire de l'ontologie MedDRA fait aussi partie du langage UMLS (Unified Medical Languages System), dont le graphe est structuré en cinq niveaux : System Organ Class, High Level Group Term, High Level Term, Preferred Term, Lowest Level Term [Mer08]. La distribution des termes dans l'ontologie MedDRA varie d'un niveau à l'autre (Table 24), et environ 37% des termes ont plus d'un parent dans l'ontologie. Dans les terminologies MedDRA et COSTART, chaque terme a un identifiant unique et possède au moins un chemin le reliant à la racine du graphe. Par exemple, le terme C0000733 : *abdomen injuries*, a comme chemin reliant à la racine C1140263.C0017178.C0947761.C0947846 et C1140263.C0947733.C0021502.C0851837.
Un exemple réel d'application de la mesure *SimDAG* concerne son utilisation pour mesurer la si-

Nombre de parents dans le rDAG	nombre de terme	pourcentage
1	12485	62.28
2	6007	29.96
3	1249	6.23
4	251	1.25
5	42	0.21
6	12	0.06
8	1	0.005

TABLE 24 – Distribution des relations parent-descendants dans la terminologie MedDRA [BBST+11].

milarité sémantique entre des termes issus de ces deux ontologies. Les matrices de similarité sont calculées avec *SimDAG*, et sont utilisées par la suite pour réaliser un clustering de termes. Nous avons utilisé 1502 termes de l'ontologie COSTART pour générer la première matrice de similarité sémantique entre toutes les paires de termes. Concernant l'ontologie MedDRA, nous avons utilisé les 1288 termes qui annotent des molécules de médicaments. Ces associations molécules-termes sont extraites de la base de donnée SIDER [14] qui est une ressource pour l'annotation des effets secondaires des médicaments et des principales molécules chimiques [CKG+08, KCL+10]. Comme dans le cas précédent avec la mesure *IntelliGO*, les étapes nécessaires pour calculer la similarité sémantique entre deux termes d'annotation sur ces ontologies supposent l'extraction du plus court chemin entre les termes et de la profondeur maximale du LCA de ces termes.
Le programme *IntelliGO* a dû être adapté pour cela (Algorithme 1). Les résultats du clustering des termes de COSTART et de MedDRA sont affichés dans la Figure 31, où le clustering hiérarchique ascendant et la visualisation par Dendroscope [15] ont été utilisés pour regrouper les termes sémantiquement similaires dans des clusters distincts, en se basant sur la similarité sémantique *SimDAG*. Nous constatons dans les deux cas, le regroupement des termes dans différents niveaux dans la hiérarchie du clustering, ce qui reflète la variation des valeurs de similarité sémantique dans les deux ontologies.

14. http ://sideeffects.embl.de/about/
15. http ://ab.inf.uni-tuebingen.de/software/dendroscope/

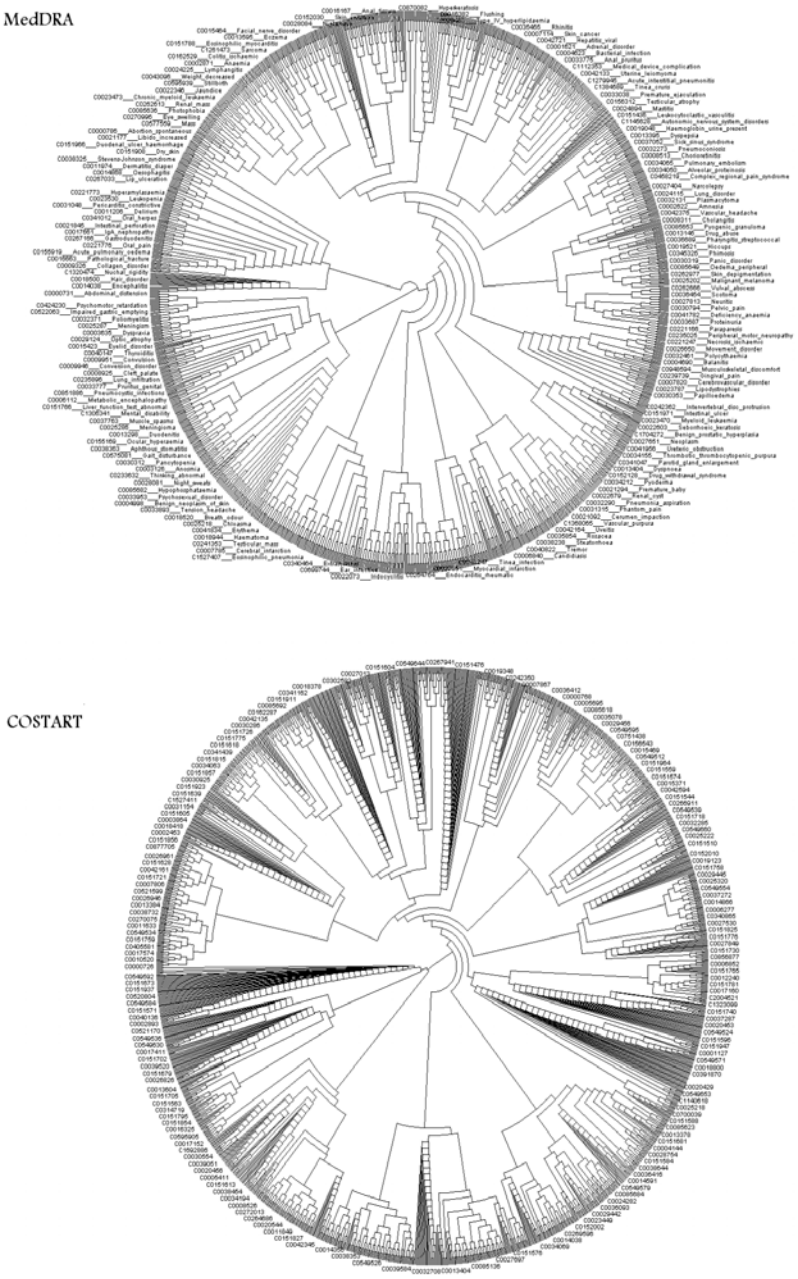


FIGURE 31 – Exemple de clustering hiérarchique et visualisation par Dendroscope de 1280 termes de l'ontologie MedDRA (haut) et 1502 termes de l'ontologie COSTART (bas), en utilisant la similarité sémantique SimDAG.

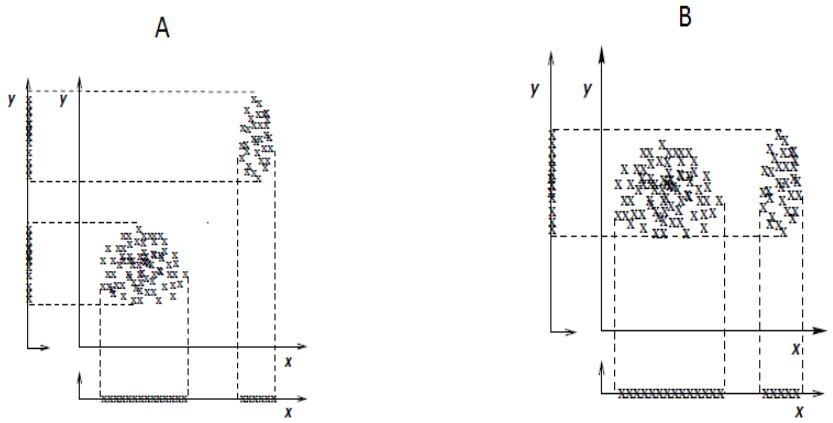


FIGURE 32 – Un exemple de redondance et de non-pertinence d'attributs de données. Dans le premier cas (A), les attributs x et y sont redondants puisque x fournit la même information que y en ce qui concerne la séparation entre les deux clusters. Dans le deuxième cas (B), l'attribut y est non pertinent car en absence de x on ne peut distinguer qu'un seul cluster [DBW04].

2 Utilisation de la mesure de similarité sémantique pour la réduction des attributs

2.1 Principe des méthodes de réduction d'attributs

2.1.1 Intérêt de la réduction d'attributs

Afin de démontrer l'intérêt de la mesure *SimDAG*, on l'a utilisée en amont d'un problème pas si courant en fouille de données et qui concerne la réduction d'attributs en utilisant les connaissances du domaine.
L'objectif général du problème de réduction d'attributs en fouille de données, est l'amélioration de l'interprétation, et la précision des modèles générés [BBST$^+$11]. En effet, plusieurs algorithmes d'apprentissage rencontrent une dégradation de performances quand des variables ou primitives non pertinentes sont présentes dans le jeu de données. Le processus de réduction d'attributs pourrait donc être considéré comme une étape importante pour maximiser la performance d'un algorithme d'extraction de connaissances à partir d'un jeu de données.

Définition: Soit ι une fonction de prédiction, et Δ un ensemble d'attributs x_1, x_2, x_3, ..., x_n. Nous définissons $X_{opt} \subseteq \Delta$, un *sous-ensemble optimal d'attributs*, telle que la performance du classifieur induit $C{=}\iota(\Delta)$ est maximale.

Le problème majeur ici réside sur le fait que X_{opt} n'est pas forcément unique. Dans la Figure 32 nous pouvons apercevoir un exemple typique montrant l'intérêt des méthodes de réduction d'attributs. Dans le premier cas (A), les attributs x et y sont redondants puisque x fournit la même information que y en ce qui concerne la séparation entre les deux clusters. Dans le deuxième cas (B), l'attribut y est non pertinent car en absence de x on ne peut distinguer qu'un seul cluster.

Plusieurs approches de réduction d'attributs ont été proposées. Les principale catégories sont présentées dans les deux sous-sections suivantes.

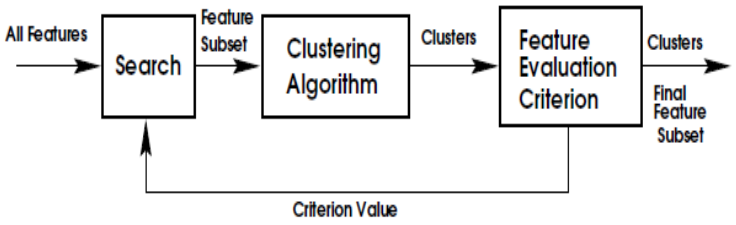


FIGURE 33 – Principe de la méthode de réduction d'attributs par agrégation (wrappers) [DBW04].

2.1.2 Réduction de données sans altération de la représentation des données : La sélection d'attributs

Dans le premier groupe de méthodes, sont rassemblées les approches de sélection de variables ou d'attributs. Elles permettent de réduire la complexité des données sans modifier les données elles-mêmes [Guy03]. Le modèle de sélection de variables (Feature selection) peut à son tour être divisé en deux sous-catégories : modèles à filtres (filters) et modèles à agrégation (wrappers). Le modèle à filtres consiste en la sélection et le pré-traitement de variables dans l'espace de recherche initial. L'évaluation des variables sélectionnées se base sur des métriques utilisant les jeux de données en entrée et est réalisée sans tenir compte de classifieurs statistiques [JKP94]. Un exemple du modèle de sélection de variables par filtre, est donné dans le cas de l'étude de corrélation entre variables, qui élimine celles qui sont redondantes et non pertinentes [WF99]. L'élagage de l'espace de recherche est souvent supervisé par des connaissances du domaines. En effet, Coulet *et al.* ont proposé d'utiliser une base de connaissances pour la sélection d'attributs dans un problème relié à la pharmagénomique [CSTB+08]. Concernant les modèles par agrégation ou encapsulation (wrappers), l'ensemble des primitives pertinentes sont sélectionnées et évaluées sur la base de leur pouvoir de prédiction en utilisant un algorithme de classification et d'apprentissage statistique [KJ97]. La classification non supervisée est généralement utilisée avec ce genre d'approche de réduction d'attributs [KSM00, DBW04]. Dans la Figure 33 nous avons une illustration de cette approche. En entrée nous avons un jeu de données initial à partir duquel, un sous-ensemble représentatif est obtenu par clustering non supervisé.

2.1.3 Réduction de données avec altération de la représentation des données : La réduction de dimension

De manière alternative aux approches précédentes, il existe des méthodes de réduction d'attributs qui modifient la représentation initiale des données en les représentant dans un espace de dimension plus petite. De telles méthodes de réduction de dimension sont aussi appelées méthodes de compression de variables ou de primitives (feature compression). Parmi ces approches, il y a l'analyse en composantes principales (ACP), qui est une méthode populaire et est couramment utilisée sur des données numériques continues. Les méthodes de clustering ont aussi été utilisées pour regrouper les attributs similaires afin d'améliorer la pertinence de la classification d'objets. Par exemple, dans la recherche d'information, des documents textuels ont des attributs qui sont binaires et correspondent aux termes annotant les documents textuels [Kyr08]. Par conséquent, le clustering des termes se base sur la comparaison de leurs distributions dans les documents [KS96]. *Solonim et al* ont proposé le paradigme d'*information bottleneck*, pour extraire

des clusters de mots qui représentent aux mieux l'information qui réside dans des documents d'un corpus [ST00]. Les auteurs remplacent par la suite, la représentation initiale des documents qui est sous forme de matrice de co-occurrence termes/document, par une représentation plus compacte basée sur la co-occurrence de clusters de termes dans les documents. En utilisant cette nouvelle représentation les termes sont regroupés en faisant référence à une mesure de similarité spécifique à un corpus. À la fin, chaque sous-ensemble d'attributs sera remplacé par une étiquette du cluster représentatif.
Dans le cas où les attributs sont issus d'un vocabulaire structuré, un moyen adéquat pour regrouper les attributs serait d'utiliser une mesure de similarité sémantique.

2.2 Clustering sémantique des attributs MedDRA en utilisant SimDAG

Afin d'illustrer l'intérêt d'utiliser la similarité sémantique *SimDAG*, dans le problème de réduction d'attributs, nous allons utiliser ici les résultats du clustering hiérarchique réalisé avec *SimDAG* sur les 1288 termes d'annotation de l'ontologie MedDRA et présentés dans la section (1.3). Le deuxième paradigme de réduction de dimension sera utilisé puisque la nature des données que nous avons ici le permet. En effet, les molécules des médicaments sont annotées par un vocabulaire du domaine décrivant les effets secondaires observés chez des patients (termes de l'ontologie MedDRA) organisé en rDAG. Avec le clustering hiérarchique, un niveau de découpe dans le dendrogramme doit être choisi pour identifier les différents clusters obtenus à ce niveau. Plusieurs méthodes ont été proposées dans la littérature à cet effet. *Kelley et al.* ont développé une technique d'optimisation du nombre de clusters, en définissant une fonction de pénalité basée sur la distance inter-cluster et la répartition des données intra-clusters [KGS96]. Quand cette valeur est minimale, les clusters résultants sont jugés bien séparés et contiennent un nombre raisonnable d'éléments avec une distance intra-cluster moyenne minimale. L'approche d'optimisation de Kelley, a été appliquée sur le clustering des 1288 termes de MedDRA (Figure 34), et la pénalité minimale est atteinte pour 112 clusters, un nombre qui paraît raisonnable selon les experts [BBST$^+$11]. Ces 112 clusters sont nommés TC (term clusters), et seront utilisés comme attributs dans une nouvelle version de l'ensemble de données initial. Un TC est affecté à un médicament si au moins un membre du TC annote le médicament dans la base de données SIDER [BBST$^+$11].
Dans le but d'étiqueter les TC avec un de ses termes le plus représentatif, nous introduisons une fonction *AvgDistTC* qui associe à tout terme MedDRA d'un TC, la moyenne des distances avec le reste des termes du même TC :

$$AvgDistTC(t_i) = \frac{1}{|TC|}\Sigma_{j=1}^{N_{TC}} dist(t_i, t_j). \tag{3}$$

Par conséquent, l'élément représentatif R de TC est le terme MedDRA t_i minimisant la valeur *AvgDistTC* :

$$R = ArgMin_t(AvgDistTC(t)). \tag{4}$$

La procédure d'étiquetage du terme représentatif a été réalisée en considérant les 112 clusters. Dans la Table 25, nous avons un exemple des distances moyennes obtenues pour chaque terme du cluster TC_{54}, donnant donc *Erythema* comme terme représentatif avec *AvgDistTC*=0,31.

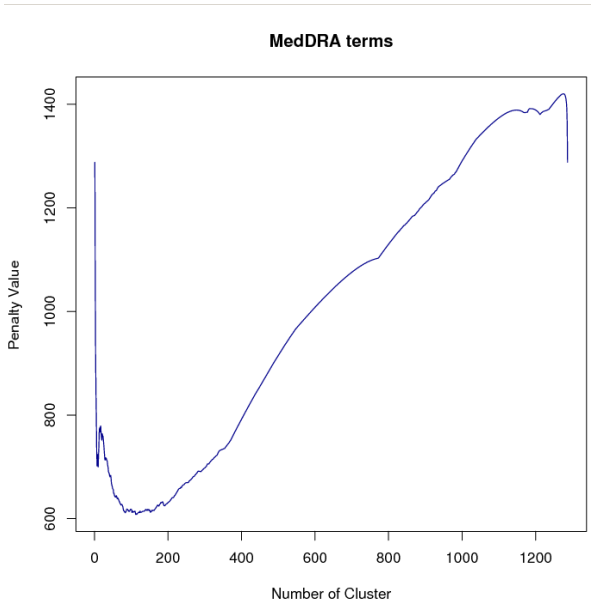


FIGURE 34 – Variation de la fonction de pénalité de la méthode d'optimisation de Kelley, en fonction du nombre de clusters de termes MedDRA [BBST+11].

2.3 Evaluation de l'impact du clustering sémantique d'attributs sur un problème de fouille de données

2.3.1 Mise en œuvre

Pour évaluer l'impact de notre stratégie de réduction de dimension, nous avons réalisé deux expériences de fouille de données. La première concerne l'extraction d'associations fréquentes d'effet secondaires partagés par des médicaments dans une certaine catégorie de médicaments. La deuxième expérience concerne la discrimination des médicaments appartenant à deux catégories selon leurs effets secondaires. Les jeux de données de test sont représentés sous forme de relations binaires (Objets × Attributs). Les effets secondaires sont représentés soit par leurs clusters représentatifs (TC), soit par les termes MedDRA eux-mêmes (sans clustering).
Dans la première expérience, nous cherchons à extraire les motifs fréquents fermés (FCI : Frequent Closed Itemsets) [YHN06], dans le but de comparer les données relatives aux médicaments annotés par les termes MedDRA dans les deux représentations, c'est-à-dire avant et après le clustering sémantique de ces termes. Dans le contexte de cette évaluation, un FCI de taille (n) et de support (s) correspond à une association de (n) termes MedDRA ou (n) clusters de termes (TC) partagés par un groupe maximal de médicaments correspondant à un pourcentage (s) du jeu de données complet. Le programme Zart disponible dans la plateforme CORON [SNK07] a été utilisé pour l'extraction des FCI, sur une machine de 1GB de mémoire et un processeur de 1,8 GHz.
Dans la deuxième expérience, nous avons utilisé l'algorithme CN2-SD [LKF+04] pour la découverte de sous-groupes dans les deux représentations des données. Cette expérience est réalisée dans le but de vérifier l'impact du clustering sémantique de termes MedDRA dans le temps du processing et des sous-groupes découverts. Ayant une population d'objets et de leurs attributs, nous sommes ici intéressés par l'extraction de sous-groupes d'objets qui sont pertinents c'est-à-dire qui sont les plus grands possibles, pour des propriétés partagées les plus spécifiques possibles, c'est-à-dire sur-représentées dans un groupe par rapport à un autre sous-groupe. Dans notre cas,

Terme (t) du cluster TC_{54}	$AvgDistTC_{54}(t)$
Decubitus ulcer	0.35
Rash	0.35
Lichen planus	0.32
Parapsoriasis	0.32
Pruritus	0.35
Psoriasis	0.37
Sunburn	0.35
Erythema	**0.31**
Pityriasis alba	0.32
Photosensitivity reaction	0.37
Rash papular	0.32
Dandruff	0.37
Lupus miliaris disseminatus faciei	0.35
Vulvovaginal pruritus	0.35

TABLE 25 – Distance moyenne calculée pour tout les termes MedDRA du cluster TC_{54}.

deux catégories de médicaments seront utilisées pour identifier des sous-groupes qui partagent des effets secondaires discriminants dans une catégorie contre l'autre. L'implémentation *Keel* de l'algorithme CN2-SD utilisée ici, est exécutée sur une machine à 8 cœurs avec 8GB de mémoire [AFH08].

2.3.2 Description des jeux de données

Une catégorie de médicaments fait référence à son application thérapeutique. Les différentes catégories de médicaments présent dans SIDER sont disponibles dans la base DRUGBANK [KLJ+11]. Pour cause de simplicité, nous avons choisi les deux grandes catégories de médicaments de cette base : Cardiovascular Agents (CA) et Anti-Infective Agents (AIA) (voir Table 26). Pour chacune des deux catégories, nous construisons deux jeux de données : en considérant d'une part la totalité des 1288 termes MedDRA, puis les 112 clusters de termes TC obtenus précédemment. Ces termes annotent tous les médicaments des deux catégories (CA) et (AIA). Le résultat est donc composé de quatre jeux de données : $CA_{allterms}$, $AIA_{allterms}$, CA_{TC} et AIA_{TC}.

Nom de la catégorie	Nombre de médicaments
Cardiovascular Agents (CA)	94
Anti-Infective Agents (AIA)	76

TABLE 26 – Nombre de médicaments dans les deux catégories choisies.

2.3.3 Extraction de motifs fréquents fermés (FCI) avec et sans clustering sémantique de termes

Le programme Zart a été exécuté sur les quatre jeux de données $CA_{allterms}$, $AIA_{allterms}$, CA_{TC} et AIA_{TC} avec un support minimal qui varie de 50 à 100%. La Table 27 résume les résultats trouvés concernant la production des FCI de chaque cas.

La première remarque que nous pouvons faire concerne le nombre de FCI qui croît en considé-

Support minimal	50%	60%	70%	80%	90%	100%
$CA_{allterms}$	386	94	41	11	1	0
CA_{TC}	5564	1379	256	62	6	0
$AIA_{allterms}$	178	41	9	2	0	0
AIA_{TC}	654	154	30	3	0	0

TABLE 27 – Nombre de FCI pour chaque jeu de données, en variant le support minimal.

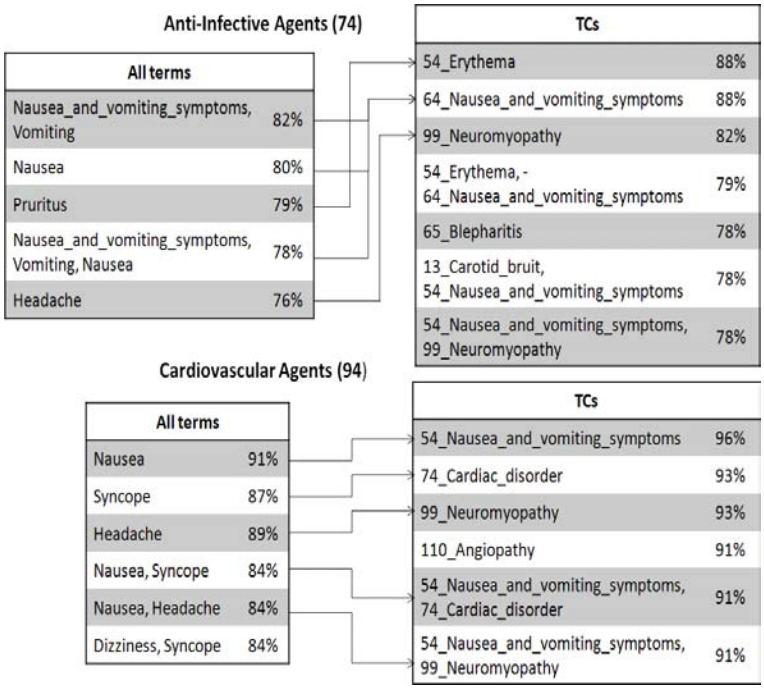


FIGURE 35 – Les cinq FCI les plus fréquents sélectionnés à partir de la liste globale des FCI générés sur chaque jeu de données. Dans le cas d'*ex-æquo* avec le cinquième élément, tous les FCI avec le même support sont affichés.

rant les TC au lieu des termes sans clustering, et avec un support minimal décroissant. Dans le cas $AIA_{allterms}$ et AIA_{TC}, cette variation est observée avec une échelle de 3 pour des supports de 50 à 70%, et de l'ordre de 1.5 pour les autres supports. Concernant les deux jeux de données $CA_{allterms}$ et CA_{TC}, la variation est de l'ordre de 14 pour un support minimal de 50 à 60%, à un ordre de 6 pour le reste des supports. Cette variation reflète clairement le rôle important que joue le clustering sémantique de termes MedDRA dans l'approche de réduction de dimension, pour la sélection de relations binaires (Objets × attributs) en agrégeant les attributs. La plupart des FCI obtenus avec la représentation (TC) correspond à des groupes de médicaments qui partagent des effets secondaires similaires présents dans un même cluster, et qui ne sont pas regroupés ensemble dans la représentation *allterms*. Il faudrait noter que le temps de traitement dans le cas de la représentation (TC) est deux fois plus long que dans le cas *allterms* du fait de la plus grande densité de la matrice de données.
L'analyse du contenu des FCI a été réalisée en les classant tout d'abord par ordre décroissant selon le support et en prenant les cinq premiers FCI (plus *ex-æquo*) qui sont dans le top de la liste. Les FCI sélectionnés sont affichés dans la Figure 35. En considérant tous les termes MedDRA sans clustering, c'est à dire la représentation (*allterms* : partie gauche de la Figure 35), les

FCI apparaissent comme très redondants. Par exemple, dans $AIA_{allterms}$ (à gauche en haut de la figure), trois des cinq FCI affichés contiennent des termes MedDRA très similaires sémantiquement : *nausea, vomiting, nausea and vomiting symptoms*. Le FCI ayant le plus grand support (82%) est même composé de deux termes synonymes. Dans le cas contraire, en considérant le clustering des termes (TCs : partie droite de la Figure 35), un FCI unique contenant l'attribut *64_ nausea and vomiting symptomes* représente le cluster des trois termes cités ci-dessus. Par conséquent, la réduction de données par clustering de termes permet d'obtenir des FCI moins redondants et offre donc des motifs potentiellement plus intéressants aux experts.
La Figure 35 illustre aussi que les supports des FCI sont plus élevés dans le cas du clustering sémantique des termes MedDRA (TC) que dans le cas des termes pris sans clustering. Une correspondance pourrait être établie entre la représentation *allterms* et la représentation TC. Cette correspondance est visible dans la Figure 35 avec des flèches. Par exemple, le FCI {54_Erythema} obtenu pour le jeu de données AIA_{TC} avec un support de 88% contient le cluster de terme *54_ Erythema* qui inclut le terme individuel *Pruritus* et qui par conséquent a été relié au FCI {Pruritus} obtenu dans le jeu de données $AIA_{allterms}$ avec un support de 79% seulement.

2.3.4 Extraction de sous-groupes avec et sans clustering sémantique de termes.

La seconde expérimentation, qui concerne l'application de notre mesure de similarité sémantique généralisée *SimDAG* à un problème de fouille de données, est représentée par l'étude de la discrimination entre des médicaments de deux catégories du point de vue de la présence ou absence d'effets secondaires. En effet, l'absence d'effets secondaires est aussi importante pour la caractérisation des médicaments. Ceci peut être pris en compte avec l'algorithme CN2-SD, puisque les entrées sont sous forme d'attributs valués. Nous avons utilisé donc cet algorithme sur l'union des jeux de données suivants auxquels est ajouté un attribut additionnel qui concerne la catégorisation : $AIA_{allterms} \cup CA_{allterms}$, et $AIA_{TC} \cup CA_{TC}$.
La première observation concerne le temps de calcul. Quand les jeux de données sont représentés avec les clusters de termes TC, le temps d'exécution prend moins de dix minutes, or dans les cas où tous les termes MedDRA sont considérés sans clustering, l'exécution ne se termine pas en six jours. Ce qui prouve que la réduction d'attributs permet l'exécution de l'algorithme CN2-SD.
La deuxième observation est faite sur les règles extraites à partir de la comparaison AIA_{TC} vs CA_{TC}. La partie gauche d'une règle est vérifiée pour un nombre de médicaments (support) parmi lesquels une fraction fait partie de la classe indiqué en partie droite (couverture). Ces valeurs (couverture/support) sont indiquées dans la partie droite de la règle. Les sous-groupes permettent d'identifier un sous-ensemble de médicaments à partir de ces sous-groupes qui partagent un profil spécifique d'effets secondaires par rapport à d'autres médicaments. Les meilleures règles, qui présentent les meilleurs pourcentages de couvertures sont affichées dans la Table 28. En résumant l'analyse sur ces résultats, nous concluons que la réduction de dimension nous a permis la découverte de sous-groupes qui était impossible dans le cas des jeux de données pris sans clustering. Une analyse approfondie de ces résultats de catégorisation des médicaments est en cours de réalisation par un expert.

3 Conclusion et Discussions

Dans ce chapitre, nous avons présenté une extension de notre mesure de similarité sémantique *IntelliGO*, initialement proposée pour l'ontologie GO. La définition étendue de la nouvelle mesure de similarité sémantique *SimDAG* nous a permis son utilisation dans une approche de

50 Angina pectoris = T AND 93 Bacteraemia = F AND 52 Ichthyosis = F AND 54 Erythema = T AND 49 Folate deficiency = F => CA (0.96/56)
31 Splenic infarction = T AND 41 Neutropenia = T AND 42 Penile discharge = F AND 77 Facial pain = T AND 79 Cachexia = T => AIA (0.88/26)

TABLE 28 – Les meilleures règles selon la couverture/support, extraites par l'algorithme CN2-SD pour les jeux de données $AIA_{TC} \cup CA_{TC}$.

réduction de dimension guidée par les connaissances du domaine. Cette méthode est basée sur le clustering sémantique des attributs en utilisant *SimDAG*. Cette approche pourrait maintenant être appliquée à divers domaines surtout dans les problèmes de fouille de données bio-médicales [LBPS, PHW$^+$07]. Nous avons testé notre approche sur 170 médicaments de la base SIDER annotés par 1288 termes du vocabulaire MedDRA, décrivant les effets secondaires de ces médicaments. En utilisant le clustering des termes fondé sur la similarité sémantique *SimDAG* nous avons réduit les 1288 attributs du jeu de données à 112 clusters de termes (TC). Par la suite nous avons adopté une représentation binaire des données réduites, où chaque TC est assigné aux médicaments concernés de la base SIDER. La représentation ignore les association multiples qui pourraient exister entre un médicament et les éléments d'un TC. Une relation multi-valuée plutôt que binaire pourrait être considérée dans ce cas. Récemment des travaux ont été proposés pour l'analyse de tels contextes formels multi-valués [MDNST08, KDKN09].
La méthode de réduction de dimension développée a été testée avec deux algorithmes de fouille de données : l'extraction des FCI et la découverte de sous-groupes. Dans le premier cas, les FCI générés à partir des TC sont moins redondants et affichent un support important par rapport à la représentation des termes sans clustering. De plus les résultats obtenus avec la réduction de dimension, semblent plus pertinents du point de vue de l'expert. Dans le cas de la découverte de sous-groupes, la réduction de dimension joue un rôle crucial. Le programme utilisé était dans l'incapacité de s'exécuter du fait de la taille de l'espace de recherche, tandis qu'avec le clustering des termes il génère des règles intéressantes capables de caractériser les effets secondaires propres à une catégorie de médicaments contre une autre. D'autres expériences pourraient être lancées pour identifier d'autres règles permettant de discriminer une catégorie donnée contre une autre.

Chapitre 6

Conclusions et perspectives

Sommaire

1 Conclusion

Tout au long du travail présenté dans ce mémoire, l'idée selon laquelle le processus de fouille de données post-génomiques devrait être guidé par les connaissances du domaine a été renforcée. En effet, la nature complexe des données transcriptomiques et leur disponibilité en masse rendent la tâche d'interprétation très difficile. Dans cette perspective, j'ai proposé et mis en œuvre un ensemble d'approches ré-utilisables, dont certaines ont déjà été mises à disposition sur Internet (http ://plateforme-mbi.loria.fr/intelligo). Ces méthodes permettent d'aider le biologiste à mieux interpréter les données transcriptomiques en prenant en compte les annotations fonctionnelles et leurs relations comme connaissance du domaine. Les méthodes de fouille de données présentées ici représentent le premier pas qui est jugé important dans un processus d'extraction de connaissances à partir de bases de données (KDD).
La première contribution est une nouvelle mesure de similarité sémantique et fonctionnelle entre gènes, appelée *IntelliGO*, qui exploite les annotations fonctionnelles et leurs relations. Ces annotations sont des termes issus d'un vocabulaire contrôlé (Gene Ontology : GO) qui est organisé en trois branches décrivant (i) le processus biologique (BP pour "Biological Process") dans lequel est impliqué le gène, (ii) la fonction moléculaire (MF pour "Molecular Function") du produit du gène, et (iii) le composant cellulaire (CC pour "Cellular Component") où se trouve localisé ce produit. Les termes du vocabulaire GO sont organisés en graphe acyclique orienté et enraciné ("rooted DAG"). La certitude avec laquelle chaque annotation est reliée à une entité biologique peut être qualifiée à l'aide de métadonnées de qualité (codes d'évidence). Les annotations du vocabulaire GO, sont largement utilisées à des fins de classification, de classement ou d'inférence de propriétés. La mesure que j'ai proposée (*IntelliGO*) prend en compte, dans une approche vectorielle inspirée de la recherche d'information, à la fois les relations de proximité dans le graphe GO et le contenu en information de ces annotations. De plus comme chaque terme d'annotation est relié à l'entité biologique qu'il annote par un code d'évidence, une pondération dépendant de

la valeur de ce code est introduite dans la mesure. La contribution essentielle de cette nouvelle mesure consiste à poser que les vecteurs normés de l'espace vectoriel et qui correspondent chacun à un terme d'annotation GO, ne sont pas tous orthogonaux deux à deux. La valeur du produit scalaire entre deux de ces vecteurs n'est donc pas nulle mais dépend de la position des termes correspondants dans la structure de GO. L'implémentation de la mesure *IntelliGO* a été conçue de façon à pouvoir prendre comme paramètres un aspect de l'ontologie GO (BP, MF, CC), un organisme particulier et une liste de poids associés aux codes d'évidence. La mesure *IntelliGO* a été validée sur des ensembles de références prédéfinis de gènes, extraits de bases de données biologiques publiques (13 KEGG Pathways pour les annotations du type BP, et 10 Pfam clans pour les annotations du type MF) pour les deux espèces (humain, levure), et comparée avec des mesures existantes. En résumé, les résultats obtenus montrent que la mesure de similarité sémantique *IntelliGO* est prometteuse, notamment dans la perspective de la classification fonctionnelle de gènes. Ainsi, la deuxième contribution concernait l'utilisation de cette mesure de similarité dans un processus de classification fonctionnelle de gènes. Deux types d'algorithme ont été utilisés dans cette approche : le clustering hiérarchique et le k-means flou. Le premier illustre la proximité entre les gènes via un arbre phylogénique, tandis que le deuxième permet un chevauchement entre clusters de gènes offrant ainsi la possibilité à un gène d'appartenir à plusieurs clusters à la fois. Au cours de l'expérimentation du k-means flou avec la mesure *IntelliGO* j'ai été amené à définir une procédure de comparaison entre les résultats de la classification fonctionnelle avec les ensembles de références utilisés précédemment pendant l'évaluation de la mesure *IntelliGO*. J'ai utilisé la méthode du F-Score, qui est un bon indicateur, souvent utilisé pour l'évaluation des algorithmes de clustering à partir de la précision et du rappel entre les clusters obtenus et les classes de référence. Les valeurs de F-Score obtenues avec la classification fonctionnelle basée sur la mesure *IntelliGO* ont été comparées avec celles obtenues avec un programme de référence de l'état de l'art.

La troisième contribution dans cette thèse concerne le regroupement, à partir d'un jeu de données transcriptomiques, des gènes partageant des niveaux d'expression proches. En effet, ayant observé que l'analyse des profils d'expression par les biologistes s'effectue souvent sous un mode différentiel, en comparant l'expression d'un gène dans deux situations données, j'ai par la suite, proposé une nouvelle approche permettant de définir des profils d'expression différentielle (PED), fondés sur l'expression différentielle des gènes entre toutes les paires de situations possibles pour le jeu de données considéré. De plus j'ai adapté une modélisation basée sur la logique floue pour affecter un gène donné à l'un ou plusieurs de ces profils d'expression. Un jeu de données constitué de 222 gènes relatifs au cancer colo-rectal a été utilisé pour évaluer l'approche d'extraction des PED et une analyse fonctionnelle en utilisant une méthode standard a été réalisée, permettant de mettre en relation des fonctions biologiques et des comportements transcriptionnels particuliers. D'autres jeux de données pourraient éventuellement être testés afin d'extraire plus de gènes qui satisferont cette hypothèse.

La dernière contribution de la thèse, concerne l'extension de la mesure de similarité sémantique *IntelliGO* à d'autres ontologies représentées sous forme de rDAG. Il s'agit d'un travail réalisé en collaboration avec E. Bresso de l'équipe Orpailleur du Loria de Nancy. À titre illustratif, nous avons appliqué la nouvelle mesure de similarité terme-terme *SimDAG* sur l'ontologie MedDRA (Medical Dictionary for Regulatory Activities) qui est utilisée comme vocabulaire contrôlé pour définir les effets secondaires de médicaments. Le calcul de la similarité *SimDAG* entre termes de l'ontologie MedDRA nous a permis de regrouper des molécules pharmaceutiques similaires du point de vue de leurs annotations par des termes de cette ontologie. De façon transversale, le clustering sémantique des termes d'annotation a été utilisé pour développer une méthode de réduction de dimension. Cette approche a été testée avec deux algorithmes de fouilles de données :

l'extraction des motifs fréquents et la découverte de sous groupes. Dans le premier cas, les motifs fréquents générés après réduction d'attributs par clustering sémantique, sont moins redondants et plus représentatifs pour l'expert. Dans le cas de la découverte de sous-groupes, la réduction de dimension joue un rôle crucial en terme d'optimisation du temps d'exécution.

2 Perspectives

2.1 Analyse transcriptomique

2.1.1 Application à d'autres jeux de données

Comme perspective de ce travail, il serait tout d'abord important de pouvoir appliquer les approches proposées à des jeux de données plus volumineux. Ce travail est en cours de réalisation dans le cadre d'un projet IBISA entre trois plateformes bioinformatiques : les biologistes du projet SIDR ("Standard-based Infrastructure with Distributed Resources", http ://sidr-dr.inist.fr/index.jsp , INIST) à Nancy qui fournissent des données transcriptomiques brutes, qui sont pré-traitées (normalisation) chez les bioinformaticiens du LBGI (IGBMC) de Strasbourg, puis finalement analysées avec mes outils (PED flous, classification fonctionnelle avec *IntelliGO*, et analyse de recouvrement) au Loria à Nancy. Les résultats sont retournés au SIDR pour enrichir les annotations du jeu de données. Ils pourront aussi servir aux bioinformaticiens du LBGI pour évaluer leurs solutions pour la normalisation des données.

2.1.2 Utilisation d'autres sources de connaissances pour guider la fouille de données

Le problème de la fouille de données dans le cas de ce travail a été essentiellement guidé par des connaissances issues de l'ontologie GO. D'autres sources de connaissances pourraient être considérées, afin d'enrichir les résultats de l'analyse des données d'expression. Par exemple, la base KEGG propose des pathways relatifs au cancer colorectal, comme le cas de la voie métabolique "*KEGG Colorectal pathway*" ou plus encore la voie métabolique "*KEGG Wnt Signaling pathway*" [JTL$^+$08, BGG$^+$07, SBVdFdP$^+$07]. Ainsi, j'ai pu vérifier que certains gènes du jeu de données avec lequel j'ai travaillé sont déjà présents dans le premier pathway.
Une autre source d'information qui pourrait être utilisée concerne les interactions entre gènes. En effet, un produit de gène peut avoir un ou plusieurs interactants, ce qui implique la possibilité d'intégration de données relationnelles entre gènes, exprimant le fait qu'il a été établi qu'un gène interagit avec d'autres gènes [BdlFM02, LLZ]. Des approches d'analyse de données multi-dimensionnelles comme l'approche d'Analyse Relationnelle de Concepts (RCA : Relational Concept Analysis) est naturellement indiquée, puisqu'elle est bien adaptée pour ce type de problème [RHNV07]. Cette approche prolonge naturellement l'analyse formelle de concepts (AFC) par le fait qu'elle enrichit les concepts formels obtenus.

2.2 Analyse de recouvrement

2.2.1 Élaboration et nettoyage de base de données biologiques

L'approche de l'étude du recouvrement et d'enrichissement des clusters fonctionnels de gènes par rapport à des ensembles de référence m'a permis de proposer, à partir des quatre jeux de données considérés, de nouvelles annotations pour certains gènes et de nouveaux gènes pour certains

ensembles de référence. Ainsi, je voudrais utiliser de nouveaux ensembles de référence sélectionnés en concertation avec des experts biologistes, avec comme objectif d'inférer de nouvelles annotations fonctionnelles pour des produits de gènes. De plus l'approche pourrait être utilisée pour aider les curateurs pour la constitution et le nettoyage de bases de données biologiques [LGKFO09].

2.3 Autres applications pour une mesure de similarité sémantique

2.3.1 Analyse formelle de concepts et clustering sémantique d'attributs : l'approche "Reverse scaling"

La mesure de similarité sémantique généralisée *SimDAG* pourrait aussi être utilisée conjointement avec l'approche d'analyse formelle de concepts (FCA). En effet, la FCA est une méthode de bi-clustering qui permet de regrouper simultanément des objets partageant un ensemble d'attributs, représentés par un concept formel. L'intérêt ici serait, dans le cas où les attributs sont très nombreux et représentés dans un vocabulaire structuré, de réaliser a priori un clustering "sémantique" entre les attributs. Ce processus permettra l'agrégation d'attributs qui vont être regroupés avec les objets qui les partagent, et par conséquent de générer des concepts formels ou des motifs fréquents facilement interprétatbles et plus pertinents vis-à-vis des besoins de l'utilisateur.

2.3.2 Prise en compte des différences entre les relations sémantiques

Dans un vocabulaire structuré, il peut y avoir plusieurs types de relations sémantiques. Pour cette raison, et puisque la similarité sémantique *SimDAG* a été proposée dans le but de son application à divers vocabulaires structurés en rDAG, je voudrais l'adapter pour tenir compte de la variabilité des relations sémantiques, en considérant de façon prioritaire les relations qui sont de type spécialisation/généralisation.

Au terme de ce travail, il apparaît que les contributions proposées tant au domaine de l'informatique qu'à celui de la bioinformatique apparaissent très prometteuses. Le caractère pluridisciplinaire de la thèse m'a contraint à appliquer les différentes approches sur des données biologiques. Cependant, il est tout à fait imaginable que ces approches puissent être appliquées à d'autres domaines d'application. Finalement, à cause de la complexité des données transcriptomiques, les outils de fouille appliqués à ce type de données devraient être à la fois robustes et satisfaisants. A ce stade, il me semble que ma contribution à ce domaine et les outils que j'ai proposés répondent à ces contraintes.

Annexe A

Quatre collections d'ensembles de référence de gènes

A.1 Liste des 13 pathways KEGG de l'espèce levure :

```
 <?xml version="1.0" encoding="UTF-8" ?>
- <Dataset xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation=
"Dataset-forwebsite.xsd" Number="1" Source_Database="KEGG
 Pathways" Release="Dec 2009">
  <Total_Set_Number>13</Total_Set_Number>
  <Total_Gene_Number>169</Total_Gene_Number>

  <Total_GO_annotations IEA="Yes">435</Total_GO_annotations>
  <Total_GO_annotations IEA="No">572</Total_GO_annotations>
  <To_Be_Tested_With>GO Biological Process Annotations</To_Be_Tested_With>
- <Pathway Species_scientificname="Saccharomyces cerevisiae"
Species_shortname="budding yeast">
  <Name KEGG_ID="Sce00562">Inositol phosphate metabolism</Name>
  <Category KEGG_ID="1">Metabolism</Category>
  <Subcategory KEGG_ID="1.1">Carbohydrate Metabolism</Subcategory>
  <Number_of_genes>15</Number_of_genes>
  <List_of_genes>851620; 851753; 851789; 851881; 850574; 856442;
  853288; 850941; 851014;  855618; 855454; 854089; 854276; 855860;
  856229;</List_of_genes>
  </Pathway>
- <Pathway Species_scientificname="Saccharomyces cerevisiae"
 Species_shortname="budding yeast">
  <Name KEGG_ID="Sce00920">Sulfur Metabolism</Name>
  <Category KEGG_ID="1">Metabolism</Category>
  <Subcategory KEGG_ID="1.2">Energy Metabolism</Subcategory>
  <Number_of_genes>13</Number_of_genes>
  <List_of_genes>850307; 850588; 850616; 852691; 852895; 853466;
  853594; 853869; 851010; 854892;
  855444; 854090; 856296;</List_of_genes>
```

```
  </Pathway>
- <Pathway Species_scientificname="Saccharomyces cerevisiae"
  Species_shortname="budding yeast">
  <Name KEGG_ID="Sce00600">Sphingolipid Metabolism</Name>
  <Category KEGG_ID="1">Metabolism</Category>
  <Subcategory KEGG_ID="1.3">Lipid Metabolism</Subcategory>
  <Number_of_genes>13</Number_of_genes>
  <List_of_genes>852481; 852568; 851634; 851888; 851891; 856386;
  853307; 853861; 853927; 850964;
  855342; 854342; 856018;</List_of_genes>
  </Pathway>
- <Pathway Species_scientificname="Saccharomyces cerevisiae"
Species_shortname="budding yeast">
  <Name KEGG_ID="Sce00300">Lysine Biosynthesis</Name>
  <Category KEGG_ID="1">Metabolism</Category>
  <Subcategory KEGG_ID="1.5">Amino Acid Metabolism</Subcategory>
  <Number_of_genes>13</Number_of_genes>
- <!--   !!! on KEGG web site, nb of genes = 11 (august 2011):
LYS5 852723 and BAN3 853386 have
been assigned to other pathways
  -->
  <List_of_genes>852412; 851425; 851346; 851736; 851820; 856778;
  852723; 852672; 854714; 854852;
   853386; 853604; 855786;</List_of_genes>
  </Pathway>
- <Pathway Species_scientificname="Saccharomyces cerevisiae"
Species_shortname="budding yeast">
  <Name KEGG_ID="Sce00410">beta-Alanine metabolism</Name>
  <Category KEGG_ID="1">Metabolism</Category>
  <Subcategory KEGG_ID="1.5">Amino Acid Metabolism</Subcategory>
  <Number_of_genes>8</Number_of_genes>
- <!--  !!! on KEGG web site, nb of genes = 11 (august 2011)
  -->
  <List_of_genes>856804; 852902; 854661; 850838; 855291; 854556;
   856044; 856182;</List_of_genes>
  </Pathway>
- <Pathway Species_scientificname="Saccharomyces cerevisiae"
Species_shortname="budding yeast">
  <Name KEGG_ID="Sce00514">Other types of O-glycan biosynthesis</Name>
  <Category KEGG_ID="1">Metabolism</Category>
  <Subcategory KEGG_ID="1.7">Glycan Biosynthesis and Metabolism
  </Subcategory>
  <Number_of_genes>13</Number_of_genes>
- <!--  on KEGG web site, nb of genes = 13 (august 2011)
  -->
  <List_of_genes>851210; 852504; 851464; 851462; 851902; 852094;
   856718; 852635; 853113; 854801;
  853608; 854266; 854499;</List_of_genes>
```

```
  </Pathway>
- <Pathway Species_scientificname="Saccharomyces cerevisiae"
Species_shortname="budding yeast">
  <Name KEGG_ID="Sce00670">One carbon pool by folate</Name>
  <Category KEGG_ID="1">Metabolism</Category>
  <Subcategory KEGG_ID="1.8">Metabolism of Cofactors and Vitamins
  </Subcategory>
  <Number_of_genes>14</Number_of_genes>
- <!--  !!! on KEGG web site, nb of genes = 15 (august 2011)
  -->
  <List_of_genes>852270; 852378; 852565; 851582; 852017; 856932;
   852752; 853118; 853955; 850715;
  850747; 855149; 854241; 854411;</List_of_genes>
  </Pathway>
- <Pathway Species_scientificname="Saccharomyces cerevisiae"
Species_shortname="budding yeast">
  <Name KEGG_ID="Sce00903" Status="dismissed by KEGG revision 11-4-2">
  Limonene and pinene degradation</Name>
- <!--  !!! problem: in up-to-date KEGG PTHWAY versions, there is no
 more sce00903, only ko00903 is found and without yeast genes
  -->
  <Category KEGG_ID="1">Metabolism</Category>
  <Subcategory KEGG_ID="1.9">Metabolism of Terpenoids and Polyketides
  </Subcategory>
  <Number_of_genes>7</Number_of_genes>
  <List_of_genes>856804; 850500; 853237; 853878; 854556; 856044;
  856163;</List_of_genes>
  </Pathway>
- <Pathway Species_scientificname="Saccharomyces cerevisiae"
Species_shortname="budding yeast">
  <Name KEGG_ID="Sce03022">Basal transcription factors</Name>
  <Category KEGG_ID="2">Genetic Information Processing</Category>
  <Subcategory KEGG_ID="2.1">Transcription</Subcategory>
  <Number_of_genes>24</Number_of_genes>
- <!--  !!! on KEGG web site, nb of genes = 32 (august 2011)
  -->
  <List_of_genes>852497; 850409; 851723; 851745; 851906; 856891;
   852839; 852766; 852888;
  853098; 853191; 853840; 853807; 853936; 850691; 854993; 854875;
   855267; 855276; 854369;
  855981; 855974; 856169; 856201;</List_of_genes>
  </Pathway>
- <Pathway Species_scientificname="Saccharomyces cerevisiae"
Species_shortname=
"budding yeast">
  <Name KEGG_ID="Sce04130">SNARE interactions in vesicular transport</Name>
  <Category KEGG_ID="2">Genetic Information Processing</Category>
  <Subcategory KEGG_ID="2.3">Folding, Sorting and Degradation</Subcategory>
```

```
  <Number_of_genes>23</Number_of_genes>
- <!--  on KEGG web site, nb of genes = 23 (august 2011)
  -->

  <List_of_genes>851219; 851203; 852079; 852109; 852780; 852660;
  852892; 856354; 854813; 853863; 853638; 850713; 850767; 850973; 855031;
   855221; 855237; 854142; 854201; 854242;
  854273; 854505; 855844;</List_of_genes>
  </Pathway>
- <Pathway Species_scientificname="Saccharomyces cerevisiae" Species_shortname
  ="budding yeast">
  <Name KEGG_ID="Sce03450" Status="Complete and OK">Non-homologous
  end-joining</Name>
  <Category KEGG_ID="2">Genetic Information Processing</Category>
  <Subcategory KEGG_ID="2.4">Replication and Repair</Subcategory>
  <Number_of_genes>10</Number_of_genes>
- <!--  on KEGG web site, nb of genes = 10 (august 2011)
  -->
  <List_of_genes>850372; 851975; 852790; 853747; 850970; 855132; 855264;
   855328; 855471;
  854166;</List_of_genes>
  </Pathway>
- <Pathway Species_scientificname="Saccharomyces cerevisiae"
Species_shortname="budding yeast">
  <Name KEGG_ID="Sce04070">Phosphatidylinositol signaling system</Name>
  <Category KEGG_ID="3">Environmental Information Processing</Category>
  <Subcategory KEGG_ID="3.2">Signal Transduction</Subcategory>
  <Number_of_genes>15</Number_of_genes>
- <!--  on KEGG web site, nb of genes = 15 (august 2011)
  -->
  <List_of_genes>852169; 852317; 852406; 851789; 851881; 850574; 856442;
   850941; 851014;
  855618; 855454; 854089; 854276; 855860; 856229;</List_of_genes>
  </Pathway>
- <Pathway Species_scientificname="Saccharomyces cerevisiae"
Species_shortname="budding yeast">
  <Name KEGG_ID="Sce04140">Regulation of autophagy</Name>
  <Category KEGG_ID="4">Cellular Processes</Category>
  <Subcategory KEGG_ID="4.1">Transport and Catabolism</Subcategory>
  <Number_of_genes>17</Number_of_genes>
- <!--  on KEGG web site, nb of genes = 17 (august 2011)
  -->
  <List_of_genes>852200; 852394; 852425; 852518; 856702; 852695;
  856576; 850684; 850941;
  851142; 855194; 855498; 855741; 855983; 855954; 856162; 856315;
  </List_of_genes>
  </Pathway>
  </Dataset>
```

A.2 Liste des 13 pathways KEGG de l'espèce humaine :

```
<?xml version="1.0" encoding="UTF-8" ?>
- <Dataset
xmlns:xsi="http://www.w3.org/2001/XMLSchemainstance"
xsi:noNamespaceSchemaLocation="Datasetcommonschema.
xsd" Number="2"
Source_Database="KEGG Pathways" Release="Dec
2009">
<Total_Set_Number>13</Total_Set_Number>
<Total_Gene_Number>275</Total_Gene_Number>
- <!--
in this dataset none of the genes belong to more
than one pathway
-->
- <!--
10 genes do not have (or did not have) any BP GO
annotation
116841,168620,1719,2822,391764,6818,6917,6919,73175
1,84265
-->
<Total_GO_annotations
IEA="Yes">620</Total_GO_annotations>
<Total_GO_annotations
IEA="No">560</Total_GO_annotations>
<To_Be_Tested_With>GO Biological Process
Annotations</To_Be_Tested_With>
- <Pathway Species_scientificname="Homo Sapiens"
Species_shortname="human">
<Name KEGG_ID="Hsa00040" Status="OK ou dismissed ou
modified">Pentose and glucuronate
interconversions</Name>
<Category KEGG_ID="1">Metabolism</Category>
<Subcategory KEGG_ID="1.1">Carbohydrate
Metabolism</Subcategory>
<Number_of_genes>25</Number_of_genes>
- <!--
!! on KEGG web site nb of genes = 33 (august 2011)
-->
<List_of_genes>10720;10941;231;2990;51181;54490;
54575;54576;54577;54578;54579;54600;54657;54658;54
659;6120;7358;7360;7363;7364;7365;7366;7367;797
99;9942;</List_of_genes>
</Pathway>
- <Pathway Species_scientificname="Homo Sapiens"
```

```
Species_shortname="human">
<Name KEGG_ID="Hsa00920">Sulfur Metabolism</Name>
<Category KEGG_ID="1">Metabolism</Category>
<Subcategory KEGG_ID="1.2">Energy
Metabolism</Subcategory>
<Number_of_genes>13</Number_of_genes>
- <!--
on KEGG web site nb of genes = 13 (august 2011)
-->
<List_of_genes>10380;166012;445329;50515;55501;6
783;6799;6817;6818;6820;6821;9060;9061;</List_of_g
enes>
</Pathway>
- <Pathway Species_scientificname="Homo Sapiens"
Species_shortname="human">
<Name KEGG_ID="Hsa00140" Status="fused with by C19
and C18-steroids in KEGG revision 11-2-10">C21-
Steroid hormone biosynthesis</Name>
- <!--
!! This pathway was probably merged with some others
as now it is called Steroid hormone biosynthesis
and contains 57 genes !!
-->
<Category KEGG_ID="1">Metabolism</Category>
<Subcategory KEGG_ID="1.3">Lipid
Metabolism</Subcategory>
<Number_of_genes>11</Number_of_genes>
- <!--
on KEGG web site nb of genes = 57 (august 2011)
-->
<List_of_genes>1109;1583;1584;1585;1586;1589;3283
;3284;3290;3291;6718;</List_of_genes>
</Pathway>
- <Pathway Species_scientificname="Homo Sapiens"
Species_shortname="human">
<Name KEGG_ID="Hsa00290">Valine, leucine and
isoleucine biosynthesis</Name>
<Category KEGG_ID="1">Metabolism</Category>
<Subcategory KEGG_ID="1.5">Amino Acid
Metabolism</Subcategory>
<Number_of_genes>11</Number_of_genes>
- <!--
on KEGG web site nb of genes = 5 (august 2011)
-->
<List_of_genes>23395;3376;51520;5160;5161;5162;55
699;57176;586;587;7407;</List_of_genes>
</Pathway>
- <Pathway Species_scientificname="Homo Sapiens"
```

```
Species_shortname="human">
<Name
KEGG_ID="Hsa00563">Glycosylphosphatidylinositol(GP
I)-anchor biosynthesis</Name>
<Category KEGG_ID="1">Metabolism</Category>
<Subcategory KEGG_ID="1.7">Glycan Biosynthesis and
Metabolism</Subcategory>
<Number_of_genes>25</Number_of_genes>
- <!--
on KEGG web site nb of genes = 25 (august 2011)
-->
<List_of_genes>10026;128869;23556;2822;284098;51
227;51604;5277;5279;5281;5283;54872;54965;55650
;80055;80235;84720;84992;8733;8818;9091;93183;9
4005;9487;9488;</List_of_genes>
</Pathway>
- <Pathway Species_scientificname="Homo Sapiens"
Species_shortname="human">
<Name KEGG_ID="Hsa00670">One carbon pool by
folate</Name>
<Category KEGG_ID="1">Metabolism</Category>
<Subcategory KEGG_ID="1.8">Metabolism of Cofactors and
Vitamins</Subcategory>
<Number_of_genes>16</Number_of_genes>
- <!--
on KEGG web site nb of genes = 18 (august 2011)
-->
<List_of_genes>10588;10797;10840;10841;123263;17
19;25902;2618;275;4522;4524;4548;471;6470;6472;7
298;</List_of_genes>
</Pathway>
- <Pathway Species_scientificname="Homo Sapiens"
Species_shortname="human">
<Name KEGG_ID="Hsa00232">Caffeine
metabolism</Name>
<Category KEGG_ID="1">Metabolism</Category>
<Subcategory KEGG_ID="1.10">Biosynthesis of Other
Secondary Metabolites</Subcategory>
<Number_of_genes>7</Number_of_genes>
- <!--
on KEGG web site nb of genes = 7 (august 2011)
-->
<List_of_genes>10;1544;1548;1549;1553;7498;9;</List
_of_genes>
</Pathway>
- <Pathway Species_scientificname="Homo Sapiens"
Species_shortname="human">
<Name KEGG_ID="Hsa03022">Basal transcription
```

```
factors</Name>
<Category KEGG_ID="2">Genetic Information
Processing</Category>
<Subcategory KEGG_ID="2.1">Transcription</Subcategory>
<Number_of_genes>38</Number_of_genes>
- <!--
!!! on KEGG web site nb of genes = 54 (august 2011)
-->
<List_of_genes>10629;11036;11037;138474;27097;29
57;2958;2959;2960;2961;2962;2963;2965;2966;2967;
2968;2969;387332;391764;51616;54457;6872;6873;6
874;6875;6877;6878;6879;6880;6881;6882;6883;688
4;6908;6917;6919;9519;9569;</List_of_genes>
</Pathway>
- <Pathway Species_scientificname="Homo Sapiens"
Species_shortname="human">
<Name KEGG_ID="Hsa03020">RNA polymerase</Name>
<Category KEGG_ID="2">Genetic Information
Processing</Category>
<Subcategory KEGG_ID="2.1">Transcription</Subcategory>
<Number_of_genes>29</Number_of_genes>
- <!--
!!! on KEGG web site nb of genes = 31 (august 2011)
-->
<List_of_genes>10621;10622;10623;11128;171568;24
6721;25885;30834;51082;51728;5430;5431;5432;543
3;5434;5435;5436;5437;5438;5439;5440;5441;54864
4;55703;64425;661;84172;84265;9533;</List_of_genes
>
</Pathway>
- <Pathway Species_scientificname="Homo Sapiens"
Species_shortname="human">
<Name KEGG_ID="Hsa04130">SNARE interactions in
vesicular transport</Name>
<Category KEGG_ID="2">Genetic Information
Processing</Category>
<Subcategory KEGG_ID="2.3">Folding, Sorting and
Degradation</Subcategory>
<Number_of_genes>38</Number_of_genes>
- <!--
!!! on KEGG web site nb of genes = 36 (august 2011)
-->
<List_of_genes>10228;10282;10490;10652;10791;112
755;116841;143187;203062;2054;23673;415117;512
72;53407;55014;55850;6616;662;6804;6809;6810;68
11;6843;6844;6845;8417;8673;8674;8675;8676;8677;
8773;9341;9342;9482;9527;9554;9570;</List_of_genes
>
```

```
</Pathway>
- <Pathway Species_scientificname="Homo Sapiens"
Species_shortname="human">
<Name KEGG_ID="Hsa03450">Non-homologous endjoining</
Name>
<Category KEGG_ID="2">Genetic Information
Processing</Category>
<Subcategory KEGG_ID="2.4">Replication and
Repair</Subcategory>
<Number_of_genes>14</Number_of_genes>
- <!--
on KEGG web site nb of genes = 14 (august 2011)
-->
<List_of_genes>10111;1791;2237;2547;27343;27434;3
981;4361;5591;64421;731751;7518;7520;79840;</List
_of_genes>
</Pathway>
- <Pathway Species_scientificname="Homo Sapiens"
Species_shortname="human">
<Name KEGG_ID="Hsa03430">Mismatch repair</Name>
<Category KEGG_ID="2">Genetic Information
Processing</Category>
<Subcategory KEGG_ID="2.4">Replication and
Repair</Subcategory>
<Number_of_genes>23</Number_of_genes>
- <!--
! on KEGG web site nb of genes = 24 (august 2011)
-->
<List_of_genes>10714;27030;2956;29935;3978;4292;4
436;4437;5111;5395;5424;5425;57804;5981;5982;59
83;5984;5985;6117;6118;6119;6742;9156;</List_of_ge
nes>
</Pathway>
- <Pathway Species_scientificname="Homo Sapiens"
Species_shortname="human">
<Name KEGG_ID="Hsa04950">Maturity onset diabetes of
the young</Name>
<Category KEGG_ID="6">Human diseases</Category>
<Subcategory KEGG_ID="6.5">Endocrine
Diseases</Subcategory>
<Number_of_genes>25</Number_of_genes>
- <!--
on KEGG web site nb of genes = 25 (august 2011)
-->
<List_of_genes>168620;2494;2645;3087;3110;3170;31
71;3172;3174;3175;3280;3375;3630;3651;389692;47
60;4821;4825;50674;5078;5080;5313;6514;6927;692
8;</List_of_genes>
```

```
</Pathway>
</Dataset>
```

A.3 Liste des 10 clans Pfam de l'espèce levure :

```
<?xml version="1.0" encoding="UTF-8" ?>
- <Dataset
xmlns:xsi="http://www.w3.org/2001/XMLSchemainstance"
xsi:noNamespaceSchemaLocation="Datasetcommonschema.
xsd" Number="3"
Source_Database="Pfam Sanger Database" Release="Oct
2009">
<Total_Set_Number>10</Total_Set_Number>
<Total_Gene_Number>118</Total_Gene_Number>
<Total_GO_annotations
IEA="Yes">366</Total_GO_annotations>
<Total_GO_annotations
IEA="No">121</Total_GO_annotations>
<To_Be_Tested_With>GO Molecular Function
Annotations</To_Be_Tested_With>
- <PfamClan Species_scientificname="Saccharomyces
cerevisiae" Species_shortname="budding yeast">
<Name PfamClan_Id="CL0328">Transmembrane di-heme
cytochrome superfamily</Name>
<Number_of_genes_considered>15</Number_of_genes_cons
idered>
<List_of_genes>850425;852734;856884;850911;853660
;854563;855797;854566;850675;854013;850736;8527
16;854583;854605;854582;</List_of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Saccharomyces
cerevisiae" Species_shortname="budding yeast">
<Name PfamClan_Id="CL0059">Six-hairpin glycosidase
superfamily</Name>
<Number_of_genes_considered>13</Number_of_genes_cons
idered>
<List_of_genes>853820;855278;856137;854708;856314
;856611;853595;850746;856470;851468;852722;8563
06;851564;</List_of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Saccharomyces
cerevisiae" Species_shortname="budding yeast">
<Name PfamClan_Id="CL0092">Actin depolymerizing
Factor</Name>
<Number_of_genes_considered>8</Number_of_genes_consid
ered>
```

```
<List_of_genes>850450;850676;852971;851635;856311
;854697;855678;856498;</List_of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Saccharomyces
cerevisiae" Species_shortname="budding yeast">
<Name PfamClan_Id="CL0099">ALDH-like
superfamily</Name>
<Number_of_genes_considered>11</Number_of_genes_cons
idered>
<List_of_genes>856044;855206;855205;854556;856804
;855137;856434;854501;856432;852291;850327;</List
_of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Saccharomyces
cerevisiae" Species_shortname="budding yeast">
<Name PfamClan_Id="CL0179">ATP-grasp
superfamily</Name>
<Number_of_genes_considered>11</Number_of_genes_cons
idered>
<List_of_genes>851126;852617;855750;853573;852507
;852818;852519;853311;853159;854295;852391;</List
_of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Saccharomyces
cerevisiae" Species_shortname="budding yeast">
<Name PfamClan_Id="CL0255">ATP synthase F0
subunit</Name>
<Number_of_genes_considered>7</Number_of_genes_consid
ered>
<List_of_genes>856027;853530;855461;854600;856435
;854509;851892;</List_of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Saccharomyces
cerevisiae" Species_shortname="budding yeast">
<Name PfamClan_Id="CL0378">ANL superfamily</Name>
<Number_of_genes_considered>10</Number_of_genes_cons
idered>
<List_of_genes>851245;850846;852329;852523;854495
;856734;854808;855288;852412;854260;</List_of_gene
s>
</PfamClan>
- <PfamClan Species_scientificname="Saccharomyces
cerevisiae" Species_shortname="budding yeast">
<Name PfamClan_Id="CL0257">N-acetyltransferase
like</Name>
<Number_of_genes_considered>18</Number_of_genes_cons
idered>
<List_of_genes>854418;855157;852228;856800;853374
```

```
;856404;856019;853167;850529;856323;856642;8561
63;856249;855091;854427;851643;853315;850892;</L
ist_of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Saccharomyces
cerevisiae" Species_shortname="budding yeast">
<Name PfamClan_Id="CL0034">Amidohydrolase
superfamily</Name>
<Number_of_genes_considered>11</Number_of_genes_cons
idered>
<List_of_genes>854845;851360;853217;855581;854973
;852587;853375;850581;852225;855304;856459;</List
_of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Saccharomyces
cerevisiae" Species_shortname="budding yeast">
<Name PfamClan_Id="CL0135">Arrestin_N-like</Name>
<Number_of_genes_considered>14</Number_of_genes_cons
idered>
<List_of_genes>854500;852837;854183;850578;852552
;853361;852960;856142;852173;853891;853393;8558
06;855318;854929;</List_of_genes>
</PfamClan>
</Dataset>
```


A.4 Liste des 10 clans Pfam de l'espèce humaine :


```
<?xml version="1.0" encoding="UTF-8" ?>
- <Dataset xmlns:xsi="http://www.w3.org/2001/XMLSchemainstance"
xsi:noNamespaceSchemaLocation="Datasetcommonschema.
xsd" Number="4" Source_Database="Pfam
Sanger Database" Release="Oct 2009">
<Total_Set_Number>10</Total_Set_Number>
<Total_Gene_Number>100</Total_Gene_Number>
<Total_GO_annotations IEA="Yes">309</Total_GO_annotations>
<Total_GO_annotations IEA="No">144</Total_GO_annotations>
<To_Be_Tested_With>GO Molecular Function
Annotations</To_Be_Tested_With>
- <PfamClan Species_scientificname="Homo sapiens"
Species_shortname="human">
<Name PfamClan_Id="CL0099">ALDH-like superfamily</Name>
<Number_of_genes_considered>18</Number_of_genes_considered>
<List_of_genes>126133;221;218;216;8854;220;219;160428;2
24;222;8659;501;64577;223;217;4329;10840;7915;</List_of
_genes>
</PfamClan>
- <PfamClan Species_scientificname="Homo sapiens"
```

```
Species_shortname="human">
<Name PfamClan_Id="CL0106">6-phosphogluconate
dehydrogenase C-terminal-like superfamily</Name>
<Number_of_genes_considered>8</Number_of_genes_considered>
<List_of_genes>1962;2819;3030;3033;5226;7358;23171;510
84;</List_of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Homo sapiens"
Species_shortname="human">
<Name PfamClan_Id="CL0417">BIR-like domains</Name>
<Number_of_genes_considered>9</Number_of_genes_considered>
<List_of_genes>329;330;331;332;4671;51530;57448;79444;1
12401;</List_of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Homo sapiens"
Species_shortname="human">
<Name PfamClan_Id="CL0165">Cache-like domain</Name>
<Number_of_genes_considered>5</Number_of_genes_considered>
<List_of_genes>9254;781;55799;93589;57685;</List_of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Homo sapiens"
Species_shortname="human">
<Name PfamClan_Id="CL0149">CoA-dependent acyltransferase
superfamily</Name>
<Number_of_genes_considered>7</Number_of_genes_considered>
<List_of_genes>54677;1375;1374;1384;1103;126129;1376;</
List_of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Homo sapiens"
Species_shortname="human">
<Name PfamClan_Id="CL0085">DHS-like NAD/FAD-binding
domain</Name>
<Number_of_genes_considered>12</Number_of_genes_considered>
<List_of_genes>23530;2108;1725;23411;23410;23408;22933
;23409;51548;51547;26061;10994;</List_of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Homo sapiens"
Species_shortname="human">
<Name PfamClan_Id="CL0076">Riboflavin synthase/Ferredoxin
reductase FAD binding domain</Name>
<Number_of_genes_considered>18</Number_of_genes_considered>
<List_of_genes>4842;4552;5447;27158;4843;4846;51706;51
700;1727;51167;606495;50507;1536;53905;50506;27035;5
0508;79400;</List_of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Homo sapiens"
Species_shortname="human">
<Name PfamClan_Id="CL0289">Folate binding domain</Name>
```

```
<Number_of_genes_considered>6</Number_of_genes_considered>
<List_of_genes>275;1757;29958;200205;55066;84705;</List_
of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Homo sapiens"
Species_shortname="human">
<Name PfamClan_Id="CL0119">Flavokinase-like nucleotidyl
transferase</Name>
<Number_of_genes_considered>7</Number_of_genes_considered>
<List_of_genes>64802;80347;23057;349565;5130;9468;5833
;</List_of_genes>
</PfamClan>
- <PfamClan Species_scientificname="Homo sapiens"
Species_shortname="human">
<Name PfamClan_Id="CL0042">Flavoprotein</Name>
<Number_of_genes_considered>10</Number_of_genes_considered>
<List_of_genes>4846;4842;4552;5447;27158;4843;55253;44
1250;1728;4835;</List_of_genes>
</PfamClan>
</Dataset>
```

Annexe B

Liste des 222 gènes différentiellement exprimés utilisés, relatifs au cancer colorectal

Probe Set ID (Affyx_ID)	Gene Symbol
212942_s_at	KIAA1199
227475_at	FOXQ1
223062_s_at	PSAT1
218182_s_at	CLDN1
228754_at	SLC6A6
204702_s_at	NFE2L3
225520_at	MTHFD1L
201195_s_at	SLC7A5
203256_at	CDH3
208394_x_at	ESM1
218145_at	TRIB3
205282_at	LRP8
230398_at	TNS4
201506_at	TGFBI
41037_at	TEAD4
1567687_at	CECR9
233550_s_at	SLC4A11
221510_s_at	GLS
225799_at	MGC4677
231008_at	UNC5CL
211399_at	FGFR2
230944_at	---
236034_at	---
211909_x_at	PTGER3
202613_at	CTPS
201479_at	DKC1
220892_s_at	PSAT1
211712_s_at	ANXA9
210937_s_at	PDX1

1560834_a_at RMST
208368_s_at BRCA2
214868_at PIWIL1
235571_at ---
1559645_at ---
229868_s_at GDF15
1559667_at ---
217382_at LOC390363 /// LOC732322
231094_s_at MTHFD1L
214121_x_at PDLIM7
221538_s_at PLXNA1
218712_at C1orf109
203367_at DUSP14
221315_s_at FGF22
205531_s_at GLS2
1562979_at ---
208560_at KCNA10
233181_at ---
1567282_at OR1J4
1554842_at SLC12A1
1560946_at ---
204775_at CHAF1B
241181_x_at ---
1569617_at OSBP2
205921_s_at SLC6A6
218726_at DKFZp762E1312
237106_at SLC11A2
216262_s_at TGIF2
208322_s_at ST3GAL1
232922_s_at C20orf59
1569095_at LOC731424
217880_at ---
229682_at MAPRE3
240047_at ---
229882_at RPS15A
217585_at NEBL
237110_at ---
1560952_at ---
1561454_at ---
242804_at C4orf15
239203_at FLJ39575
223570_at MCM10
236853_at C13orf16
216488_s_at ATP11A
233104_at C20orf119
1563027_at ---
205937_at CGREF1
239438_at RAPGEF6

230811_at C16orf55
1553976_a_at RP11-529I10.4
204558_at RAD54L
240909_at ---
212789_at NCAPD3
223813_at PRND
1556459_at ---
1566924_at ---
209544_at RIPK2
211702_s_at USP32
226569_s_at CHTF18
226552_at IER5L
230021_at C15orf42
239700_at ---
244750_at ---
241126_at ---
1562219_at FLJ41649
1558046_x_at LOC441528
1563391_at ---
215576_at ---
235992_s_at LOC606495
207917_at ---
212210_at INTS1
241746_at CUL7
231564_at MANEAL
1553927_at C7orf33
208401_s_at GLP1R
241475_at BREA2
241143_at ---
1560560_at ---
239617_at ---
210316_at FLT4
1555550_at LGICZ1
235038_at KRR1
1562906_at LOC340069
241896_at ---
237407_at HS1BP3
205254_x_at TCF7
1563043_at LOC285375
1558381_a_at ---
213507_s_at KPNB1
244886_at LOC389641
1554914_at PLA2G4D
1564122_at LOC283875
217366_at CTNNA1
236425_at ---
1556958_at ---
1552796_a_at SIM1

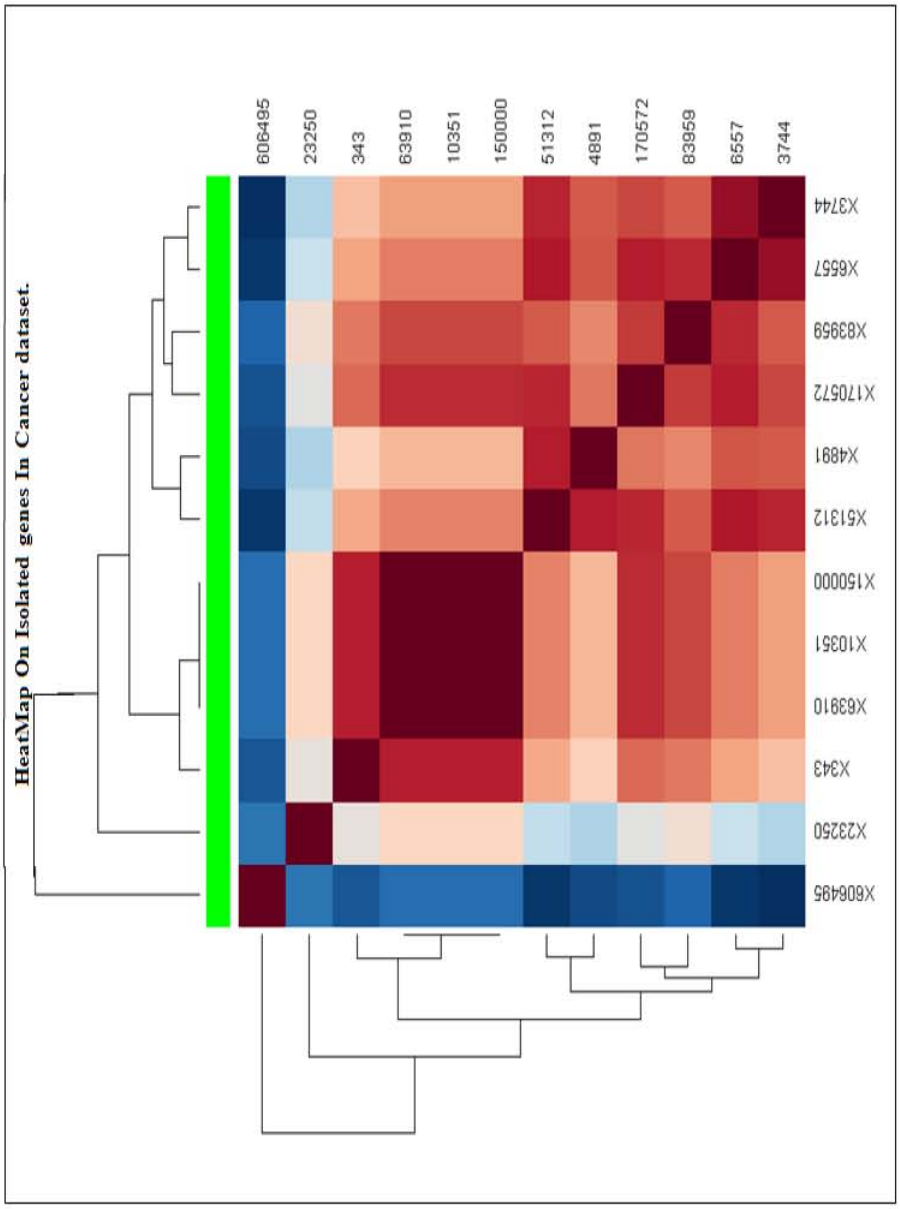
244440_at ---
210979_at ASAH1
210531_at NR2C1
1564887_at ---
205598_at TRAIP
219490_s_at DCLRE1B
218023_s_at FAM53C
209176_at SEC23IP
231757_at TAS2R5
1559159_at CEP68
222900_at ---
234434_at ---
240396_at IL20RA
243750_x_at C21orf70
237843_at ---
240547_at KIF27
215834_x_at SCARB1
243989_at ---
217142_at LOC442215
237191_x_at ---
206258_at ST8SIA5
1554530_at VSTM2
1552582_at ABCC13
230117_at LOC285878
219368_at NAP1L2
230788_at GCNT2
231925_at ---
1558747_at SMCHD1
213309_at PLCL2
213716_s_at SECTM1
211737_x_at PTN
225575_at LIFR
228885_at MAMDC2
210198_s_at PLP1
229839_at SCARA5
209735_at ABCG2
209687_at CXCL12
AFFX-LysX-3_at ---
204719_at ABCA8
206784_at AQP8
34260_at TELO2
1556299_s_at FLJ40142
237211_x_at MORN3
1564511_a_at FSTL4
213099_at ANGEL1
1556162_at IGSF3
1562020_s_at NT5DC4
211445_x_at NACA3P /// NACAP1

1563108_at LRRC62
239774_at ---
1554404_a_at ---
233352_at ---
201265_at ---
233475_at MGC32805
207466_at GAL
2028_s_at E2F1
1554735_a_at TRIM7
232917_at ARMC10
231002_s_at RABEP1
231273_x_at FRZB
208471_at HPR
236968_at C12orf12
1561976_at C1orf167
216569_at ---
1553041_at HTR3C
1556075_at ---
1552402_at CALML6
222554_s_at NOL6
1557444_at TREML3
239408_at ---
231811_at CXYorf1 /// FAM39B /// FAM39DP /// FLJ00038 /// LOC653635
204788_s_at PPOX
1560213_at LOC642032
1554837_a_at CYP4A11
240962_at ---
241165_at ---
1562426_a_at FARP1
1559728_at ZBTB40
236530_at HS1BP3
242567_at ABTB1
216600_x_at ---
229786_at ---
237618_at ---
231156_at ---
64488_at ---
208145_at ---
31835_at HRG
242597_at ---
1553340_s_at AMAC1 /// AMAC1L2
215697_at ---
244058_at C10orf72
1560286_s_at ---
1560279_a_at LOC221122
203731_s_at ZKSCAN5
207518_at DGKE
225920_at LOC148413

241127_at ---

Annexe C

Analyse du cluster fonctionnel isolé des gènes relatifs au cancer colorectal



NCBI Id	*Profil*
23250	P1
10351	P14
150000	P14
343	P14
3744	P2
6557	P2
83959	P2
4891	P2, P20
606495	P2, P3
170572	P2, P3
51312	P20
63910	P3

FIGURE C.1 – Détail de la heatmap du cluster fonctionnel isolé

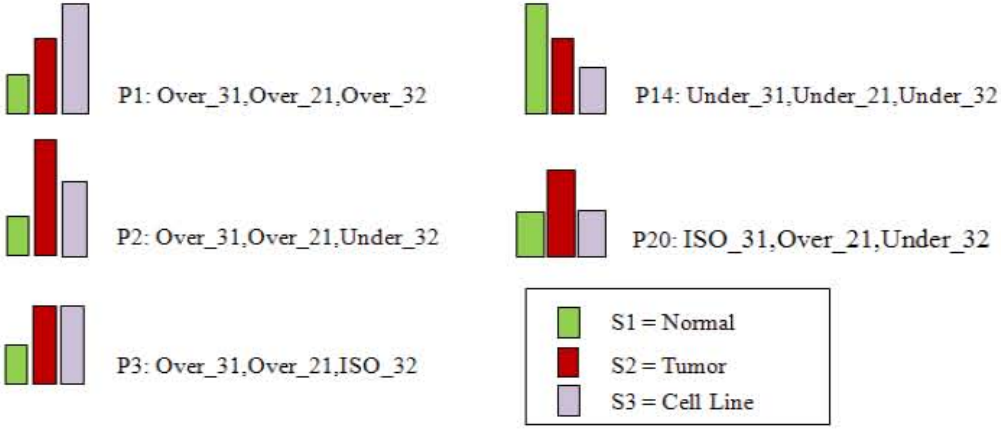
Analysis of the genes present in the isolated functional cluster in the set of 222 genes
(Genes are ordered by DEP profile)

Infos retrieved from GENE-NCBI database , Oct 13, 2010.

Name	GeneID	Full Name	GO terms (BP)	DEP-Profile	Ref to colorectal cancer colorectal or disease or drug	OMIM relatedID	Pathway
ATP11A	23250	ATPase, class VI, type 11A	ATP biosynthetic process (IEA) phospholipid transport (IEA)	P1	(1) ATP11A is a novel predictive marker for metachronous metastasis of colorectal cancer. Miyoshi N, et al. Oncol Rep, 2010 Feb. PMID 20043114.	*605868	
ABCA8	10351	ATP-binding cassette, sub-family A (ABC1), member 8	Transport (IDA)	P14	(2) Functional analysis of ABCA8, a new drug transporter. Tsuruoka S, et al. Biochem Biophys Res Commun, 2002 Oct 18. PMID 12379217.	*612505	KEGG pathway: ABC transporters 02010
ABCC13	150000	ATP-binding cassette, sub-family C (CFTR/MRP), member 13, pseudogene	Transmembrane transport (IEA)	P14		*608835	
AQP8	343	Aquaporin 8	canalicular bile acid transport (IEA) cellular response to cAMP (IEP) transport (TAS) water transport (IEA)	P14	Aquaporin 8 is expressed in colonic epithelial cells but was not detected in colorectal tumors. (3) Differential expression of Aquaporin 8 in human colonic epithelial cells and colorectal tumors Heléne Fischer, Roger Stenling, Carlos Rubio, and Annika Lindblom, BMC Physiol. 2001; 1: 1.	*603750	
KCNA10	3744	potassium voltage-gated channel, shaker-related subfamily, member 10	ion transport (IEA) potassium ion transport (IEA) transmembrane transport (IEA)	P2		*602420	
SLC12A1	6557	solute carrier family 12	ion transport (IEA)	P2		*600839	Reactome

		(sodium/potassium/chloride transporters), member 1	potassium ion transport (IEA) sodium ion transport (IEA) transmembrane transport (IEA) transport (TAS)			#601678	Event:Transmembrane transport of small molecules REACT_15518
SLC4A11	83959	solute carrier family 4, sodium borate transporter, member 11	anion transport (IEA) bicarbonate transport (IDA) boron transport (IDA) cellular cation homeostasis (IDA) fluid transport (ISS) phosphoenolpyruvate-dependent sugar phosphotransferase system (IEA) proton transport (IDA) sodium ion transport (IDA)	P2	congenital hereditary endothelial dystrophy	*610206 #217700 #217400 #613268	
SLC11A2	4891	solute carrier family 11 (proton-coupled divalent metal ion transporters), member 2	activation of caspase activity (IDA) cadmium ion transmembrane transport (IDA) cellular response to oxidative stress (IDA) cobalt ion transport (IDA) copper ion transport (IDA) etc. (25 GO terms !)	P2, P20		*600523 #206100	KEGG pathway: Lysosome 04142
CYB5RL	606495	cytochrome b5 reductase-like	oxidation reduction (IEA)	P2, P3	Hors cluster		
HTR3C	170572	5-hydroxytryptamine (serotonin) receptor 3, family member C	ion transport (IEA)	P2, P3		*610121	
SLC25A37	51312	solute carrier family 25,	ion transport (IEA)	P20		*610387	

		member 37	mitochondrial iron ion transport (IEA) transmembrane transport (IEA)				
SLC17A9	63910	solute carrier family 17, member 9	Exocytosis (IEA) transmembrane transport (IEA)	P3		*612107	
= 12 gènes							



Annexe D

Résultats d'enrichissement des gènes qui sont dans un PED et un cluster obtenu avec la classification fonctionnelle utilisant IntelliGO.

Cluster 1		Cluster 1 ⊓ Profil 3	
Terme GO	P_value	Terme GO	P_value
•NADH oxidation	• 5,08E-06	•regulation of transcription, DNA-dependent	•9,95E-04
•vacuolar proton-transporting V-type ATPase complex assembly	• 8,78E-04		
•glutamine catabolic process	• 1,62E-03	•NADH oxidation	•2,98E-04
•apical junctior assembly	• 3,20E-03	•vacuolar proton-transporting V-type ATPase complex assembly	•3,97E-03
•fatty acid alpha-oxidation	• 3,20E-03		
•L-serine biosynthetic process	• 3,20E-03		
•pyridoxine bicsynthetic process	• 6,38E-03	•fatty acid alpha-oxidation	•5,86E-03
•regulation of MAP kinase activity	• 6,74E-03	•fructose 1,6-bisphosphate metabolic process	•8,89E-03
•fructose 1,6-bisphosphate metabolic process	• 9,54E-03		
•pyrimidine nucleotide biosynthetic process	• 9,54E-03	•oxidation reduction	•1,18E-02
•glycosphingolipid biosynthetic process	• 9,54E-03	•positive regulation of ATPase activity	•1,38E-02
•folic acid and derivative biosynthetic process	• 1,27E-02	•heme biosynthetic process	•3,38E-02
•cellular copper ion homeostasis	• 1,72E-02	•glutamine metabolic process	•3,56E-02
•positive regulation of ATPase activity	• 2,20E-02	•amino acid metabolic process	•7,23E-02
•pseudouridine synthesis	• 2,20E-02	•glycolysis	
•telomere maintenance via telomerase	• 2,38E-02	•rRNA processing	
•phospholipid catabolic process	• 2,50E-02		
•ceramide metabolic process	• 2,79E-02		
•glycosaminoglycan biosynthetic process	• 2,81E-02		
•CTP biosynthetic process	• 2,81E-02		
•heme biosynthetic process	• 2,81E-02		
•chromosome organization	• 2,81E-02		
•glutamine metabolic process	• 2,87E-02		
•amino acid bicsynthetic process	• 3,11E-02		
•cellular iron ion homeostasis	• 3,11E-02		
•negative regulation of protein kinase activity	• 3,41E-02		
•one-carbon compound metabolic process	• 3,41E-02		
•amino acid metabolic process	• 3,71E-02		
•glycolysis	• 4,30E-02		
•protein homooligomerization	• 6,59E-02		
•cell-cell adhesion	• 6,59E-02		
•RNA processirg	• 6,80E-02		
•steroid metabolic process	• 8,23E-02		
•fatty acid metabolic process	• 8,50E-02		
•nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	• 8,77E-02		
•protein amino acid glycosylation			

FIGURE D.1 – Analyse d'enrichissement des termes GO des gènes présents dans le cluster 1 avant et après le recouvrement par le profil le mieux apparié (cluster1, profil 3).

Cluster 2		Cluster 2 ⊓ Profil 1	
Terme GO	P_value	Terme GO	P_value
•insulin secretion	•8,80E-06	•chromosome organization	•9,5528E-05
•vascular endothelial growth factor receptor signaling pathway	•5,56E-04	•double-strand break repair via homologous recombination	•1,1012E-04
•glucose metabolic process	•7,02E-04		•3,3740E-04
•positive regulation of apoptosis	•1,28E-03	•response to estrogen stimulus	•9,6584E-04
•chromosome organization	•1,76E-03	•DNA recombination	•1,0664E-03
•integrin-mediated signaling pathway	•1,78E-03	•mitotic metaphase/anaphase transition	•1,0664E-03
•endocrine pancreas development	•2,02E-03	•response to genistein	•1,0664E-03
•double-strand break repair via homologous recombination	•2,02E-03	•cell proliferation	•2,1307E-03
•cell maturation	•3,23E-03	•negative regulation of mammary gland epithelial cell proliferation	•1,1474E-03
•positive regulation of megakaryocyte differentiation	•4,55E-03		•2,0637E-03
•mitotic metaphase/anaphase transition	•4,55E-03	•positive regulation of apoptosis	•2,1307E-03
•synaptic vesicle membrane organization	•4,55E-03	•regulation of glucocorticoid metabolic process	•2,1307E-03
•hormone secretion	•4,55E-03	•cytokinesis during cell cycle	•2,2097E-03
•negative regulation of systemic arterial blood pressure	•4,55E-03	•chordate embryonic development	•3,1928E-03
•regulation of lung blood pressure	•4,55E-03	•response to drug	•3,1928E-03
•long-chain fatty acid biosynthetic process	•4,55E-03	•growth hormone secretion	•3,1928E-03
•response to genistein	•5,96E-03	•negative regulation of lymphocyte proliferation	•4,2529E-03
•negative regulation of mammary gland epithelial cell proliferation	•5,96E-03	•replication fork protection	•4,2529E-03
•fibroblast growth factor receptor signaling pathway	•5,96E-03	•regulation of S phase of mitotic cell cycle	•5,3108E-03
•response to estrogen stimulus	•6,01E-03	•T cell proliferation	•5,3108E-03
•vasculogenesis	•8,56E-03	•inner cell mass cell proliferation	•5,3108E-03
•angiogenesis	•9,06E-03	•DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator	•5,3108E-03
•response to stress	•9,06E-03		•5,3108E-03
•negative regulation of vasoconstriction	•9,06E-03	•L-amino acid transport	•5,3108E-03
•myelination in the central nervous system	•9,06E-03	•female gonad development	•6,3666E-03
•actin nucleation	•9,06E-03	•response to UV-C	•7,4203E-03
•cytokinesis during cell cycle	•9,06E-03	•centrosome duplication	•7,4203E-03
•regulation of glucocorticoid metabolic process	•9,06E-03	•oocyte maturation	
•positive regulation of secretion	•9,06E-03	•positive regulation of mitotic cell cycle	
•chordate embryonic development	•9,06E-03	•neutral amino acid transport	
•stress-activated MAPK cascade	•1,35E-02		
•regulation of actin filament length	•1,35E-02		
•DNA replication-dependent nucleosome assembly			
•growth hormone secretion			
•replication fork protection			

FIGURE D.2 – Analyse d'enrichissement des termes GO des gènes présents dans le cluster 2 avant, et après le recouvrement par le profil le mieux apparié (cluster 2, profil 1).

Cluster 2 ⊓ Profil 2		Cluster 2 ⊓ Profil 20	
Terme GO	P_value	Terme GO	P_value
•cell differentiation	•7,35E-05	•cell differentiation	•5,97E-05
•vascular endothelial growth factor receptor signaling pathway	•1,06E-04	•multicellular organismal development	•9,58E-05
•angiogenesis	•5,83E-04	•insulin secretion	•1,42E-04
•multicellular organismal development	•1,10E-03	•synaptic vesicle membrane organization	•1,21E-03
•fibroblast growth factor receptor signaling pathway	•1,19E-03	•hormone secretion	•1,21E-03
•positive regulation of megakaryocyte differentiation	•1,99E-03	•negative regulation of systemic arterial blood pressure	•1,21E-03
•synaptic vesicle membrane organization	•1,99E-03	•regulation of lung blood pressure	•1,21E-03
•negative regulation of systemic arterial blood pressure	•1,99E-03	•signal transduction	•1,34E-03
•regulation of lung blood pressure	•1,99E-03	•negative regulation of vasoconstriction	•2,41E-03
•positive regulation of secretion	•3,97E-03	•stress-activated MAPK cascade	•2,41E-03
•actin nucleation	•3,97E-03	•regulation of actin filament length	•2,41E-03
•negative regulation of vasoconstriction	•3,97E-03	•exocrine pancreas development	•3,62E-03
•regulation of actin filament length	•4,20E-03	•norepinephrine metabolic process	•3,62E-03
•response to stimulus	•5,14E-03	•angiogenesis	•4,39E-03
•visual perception	•5,95E-03	•regulation of transcription from RNA polymerase II promoter in response to oxidative stress	•4,82E-03
•norepinephrine metabolic process	•7,92E-03	•surfactant homeostasis	•4,82E-03
•actin filament severing	•7,92E-03	•synaptic vesicle maturation	•6,02E-03
•surfactant homeostasis	•7,92E-03	•negative regulation of fibrinolysis	•6,02E-03
•regulation of transcription from RNA polymerase II promoter in response to oxidative stress	•7,92E-03	•morphogenesis of embryonic epithelium	•8,40E-03
•calcium ion-dependent exocytosis	•9,88E-03	•transmembrane receptor protein serine/threonine kinase signaling pathway	•8,40E-03
•epithelial cell development	•1,18E-02	•regulation of heart rate	•8,40E-03
•synaptic vesicle maturation	•1,38E-02	•cAMP-mediated signaling	•9,59E-03
•chemosensory behavior	•1,38E-02	•vascular endothelial growth factor receptor signaling pathway	•1,31E-02
•regulation of heart rate	•1,38E-02	•alveolus development	•1,43E-02
•transmembrane receptor protein serine/threonine kinase signaling pathway	•1,57E-02	•nitric oxide mediated signal transduction	•1,67E-02
•regulation of cell-matrix adhesion	•1,76E-02	•blood vessel remodeling	•1,67E-02
•transcription, DNA-dependent	•1,76E-02	•embryonic placenta development	•1,78E-02
•positive regulation of cell cycle	•1,96E-02	•mitochondrion organization	•1,78E-02
•regulation of mitotic metaphase/anaphase transition	•2,15E-02	•regulation of GTPase activity	•1,78E-02
•endoderm development	•2,53E-02	•endocrine pancreas development	•1,90E-02
•negative regulation of transforming growth factor beta receptor signaling pathway	•2,72E-02	•positive regulation of bone mineralization	•1,90E-02
•alveolus development		•erythrocyte differentiation	
•cilium assembly		•mesoderm formation	
•blood vessel remodeling.			

FIGURE D.3 – Analyse d'enrichissement des termes GO des gènes présents dans l'intersection des paires (cluster 2, profil 20) et (cluster 2, profil 2).

Cluster 3		Cluster 3 ∏ Profil 14	
Terme GO	P_value	Terme GO	P_value
•ion transport	•8,90E-06	•Water transport	•2,08E-05
•Water transport	•1,22E-04	•transport	•2,56E-03
•iron ion transport	•5,04E-04		
•beta-alanine transport	•1,35E-03		
•taurine transport	•1,35E-03		
•cellular cation homeostasis	•1,35E-03		
•boron transport	•1,35E-03		
•ferrous iron transport	•1,35E-03		
•ribosomal protein import into nucleus	•4,04E-03		
•response to iron ion	•5,38E-03		
•bicarbonate transport	•6,72E-03		
•cobalt ion transport	•6,72E-03		
•response to zinc ion	•6,72E-03		
•sodium ion transport	•1,17E-02		
•potassium ion transport	•1,59E-02		
•transport	•1,60E-02		
•NLS-bearing substrate import into nucleus	•1,73E-02		
•protein import into nucleus, translocation	•2,12E-02		
•phospholipid transport	•2,12E-02		
•protein import into nucleus, docking	•2,25E-02		
•anion transport	•2,64E-02		
•cellular iron ion homeostasis	•2,89E-02		
•amino acid metabolic process	•4,53E-02		
•neurotransmitter transport	•5,02E-02		
•protein localization	•5,02E-02		
•chloride transport	•5,98E-02		
•ATP biosynthetic process	•6,10E-02		
•proton transport	•6,93E-02		
•response to drug	•8,32E-02		

FIGURE D.4 – Analyse d'enrichissement des termes GO des gènes présents dans le cluster 3 avant, et après le recouvrement par le profil le mieux apparié (cluster 3, profil 14).

Annexe E

Captures d'écran de l'interface web de l'outil IntelliGO

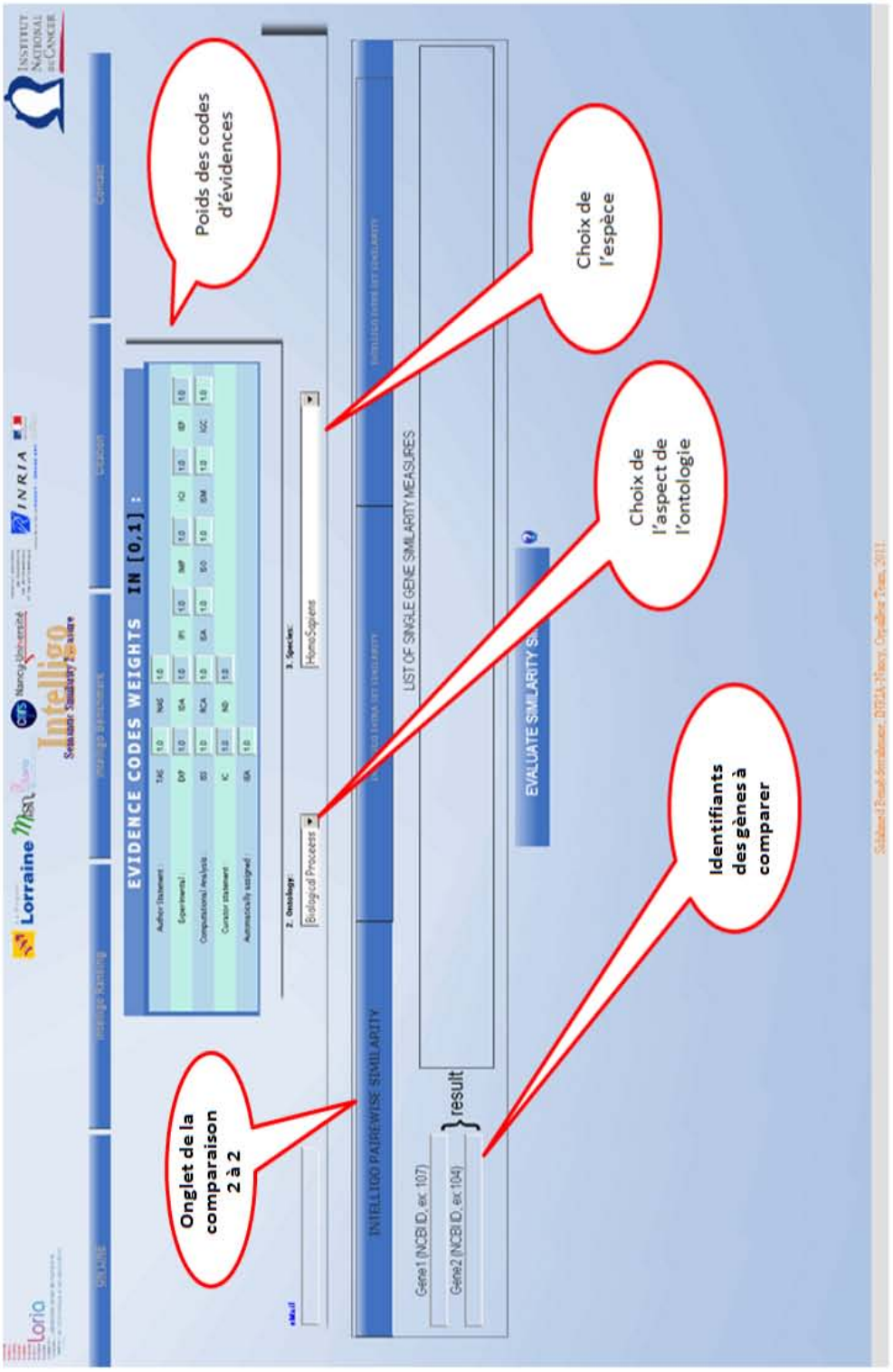


FIGURE E.1 – Capture d'écran de l'interface web de l'outil IntelliGO : page principale pour le choix des paramètres.

FIGURE E.2 – Exemple de calcul de la similarité Intra-set pour une liste de gènes.

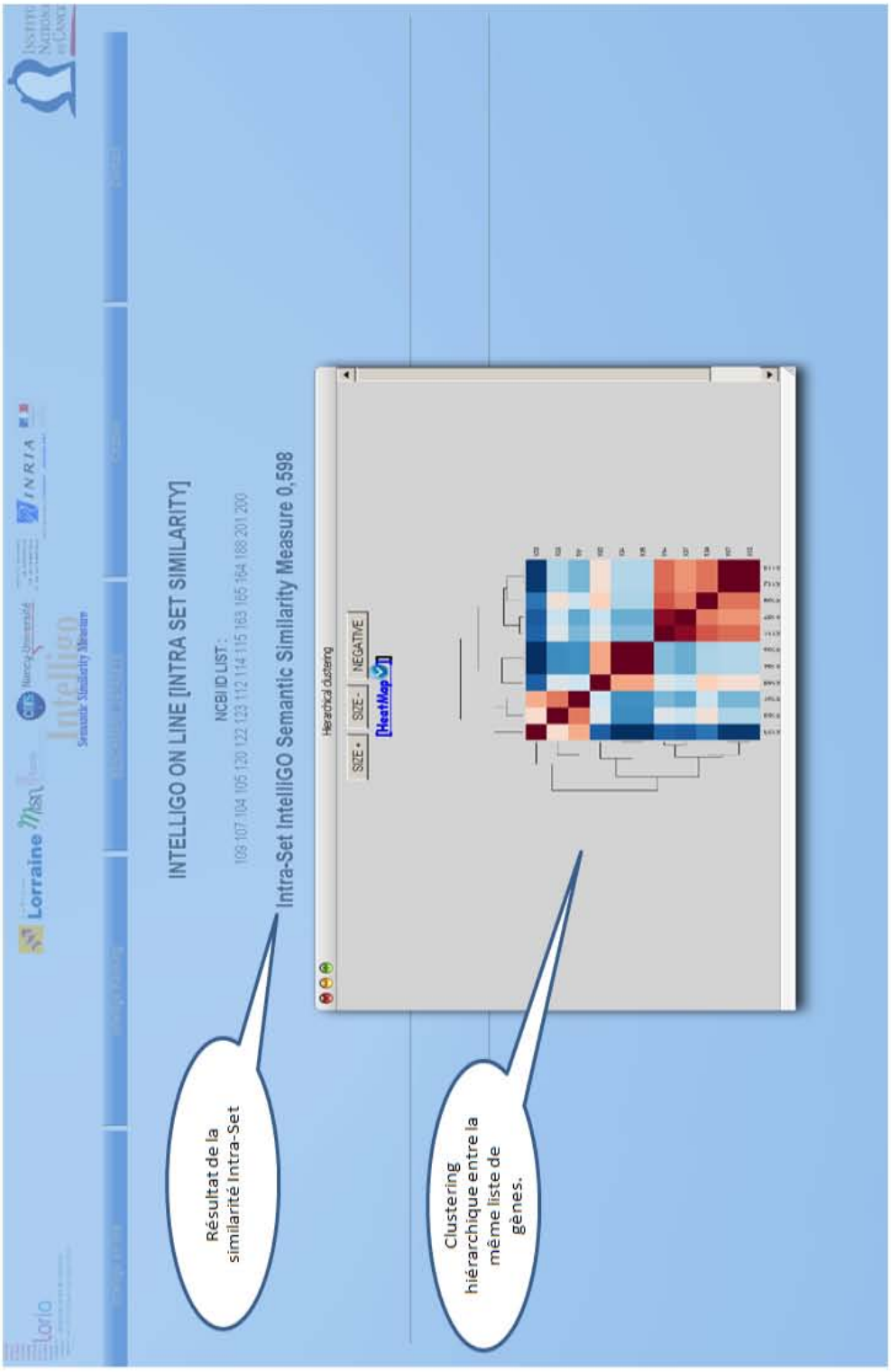


FIGURE E.3 – Résultat du clustering hiérarchique et visualisation par heatmap.

FIGURE E.4 – Résultat de la classification fonctionnelle floue.

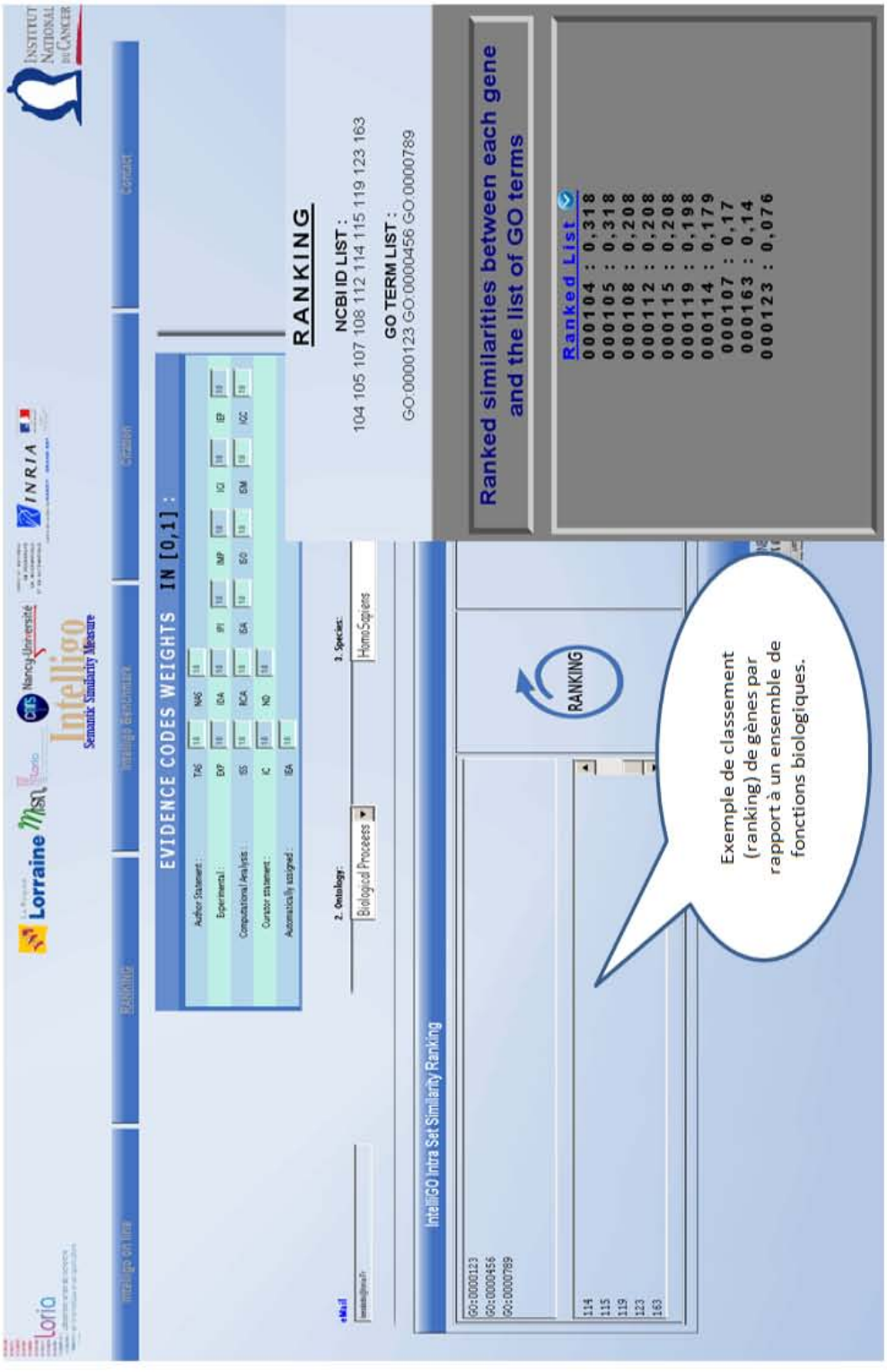


FIGURE E.5 – Exemple de l'utilisation d'IntelliGO dans un problème de classement (ranking) de gènes. Au départ une liste de gènes est donnée en paramètre avec une liste de termes GO (fonctions) d'intérêts. La similarité sémantique entre les gènes et les termes GO, est calculée et les gènes sont classés par ordre décroissant par rapport aux valeurs de similarité obtenues. Les gènes se trouvant dans le top de la liste triée sont fortement similaires aux fonctions d'intérêt.

Annexe F

Code source pour le calcul de similarité sémantique entre des objets annotés par des termes d'annotation

```
/*
SEM SIM PROJECT
Free For Use, just contact the author.
Author: Sidahmed Benabderrahmane,
sidahmed.benabderrahmane@gmail.com
*/
import java.io.IOException;
import java.io.PrintWriter;
import java.util.StringTokenizer;
import java.util.ArrayList;
import java.sql.*;
import java.io.*;
import java.util.*;
import org.ini4j.Ini;
import org.ini4j.IniPreferences;
import java.util.Map ;
import java.text.*;
import java.lang.Math;
import java.lang.Boolean;
public class SemanticSIMILARITY{

/**
* @param args

*/
public static void main(String[] args) {
// TODO Auto-generated method stub
// TODO code application logic here
Ini ini = new Ini();
Ini inisim = new Ini();
```

```
ArrayList<String> drug_list1 =new ArrayList<String>();drug_list1.clear();
ArrayList<String> drug_list2 =new ArrayList<String>();drug_list2.clear();
String path=args[0].toString().trim();
System.out.print(args[0].toString().trim());

//--fichier argument 1, liste de termes
try{
BufferedReader buff = new BufferedReader(new FileReader(path));

try {
String line;
while ((line = buff.readLine()) != null) {
drug_list1.add(line.toString().trim());
}
} finally {

buff.close();
}
} catch (IOException ioe) {

System.out.println("Erreur --" + ioe.toString());

}
//----2° fichier
path=args[1].toString().trim();
System.out.print(args[1].toString().trim());
try{
//Création du flux bufférisé sur un FileReader, immédiatement suivi par un
//try/finally, ce qui permet de ne fermer le flux QUE s'il le reader
//est correctement instancié (évite les NullPointerException)
BufferedReader buff = new BufferedReader(new FileReader(path));

try {
String line;
//--Lecture du fichier ligne par ligne. Cette boucle se termine
//--quand la méthode retourne la valeur null.
while ((line = buff.readLine()) != null) {
drug_list2.add(line.toString().trim());
//System.out.println(line);
//faites ici votre traitement
}
} finally {
//dans tous les cas, on ferme nos flux
buff.close();
}
} catch (IOException ioe) {
//--erreur de fermeture des flux
System.out.println("Erreur --" + ioe.toString());
```

```
}
path=args[2].toString().trim();
System.out.print(args[2].toString().trim());
try{
//Création du flux bufférisé sur un FileReader, immédiatement suivi par un
//try/finally, ce qui permet de ne fermer le flux QUE s'il le reader
//est correctement instancié (évite les NullPointerException)

ini.load(new FileReader(path));

} catch (IOException ioe) {
//--erreur de fermeture des flux
System.out.println("Erreur --" + ioe.toString());
}
//------
path=args[3].toString().trim();
System.out.print(args[3].toString().trim());
try{
//Création du flux bufférisé sur un FileReader, immédiatement suivi par un
//try/finally, ce qui permet de ne fermer le flux QUE s'il le reader
//est correctement instancié (évite les NullPointerException)

inisim.load(new FileReader(path));

} catch (IOException ioe) {
//--erreur de fermeture des flux
System.out.println("Erreur --" + ioe.toString());

}
//-------------------------------------------------
Connection conn = null;
String url ="jdbc:mysql://gbmserv.loria.fr:3306/"; /
String dbName database";
String driver ="com.mysql.jdbc.Driver";// "org.postgresql.Driver";
String userName ="user";
String password ="passwd";

Statement state;
ArrayList<String> terme_Liste1 =new ArrayList<String>();
terme_Liste1.clear();
ArrayList<String> terme_Liste2 =new ArrayList<String>();
terme_Liste2.clear();
ArrayList<String> IAF1 =new ArrayList<String>();
IAF1.clear();
ArrayList<String> IAF2 =new ArrayList<String>();
IAF2.clear();
```

```
@SuppressWarnings("unused")
String terms_of_drug= "SELECT *  FROM ‘sider‘ INNER JOIN ‘side_effect_meddra_synonymous‘
ON ‘sider‘.‘co_id‘ = ‘side_effect_meddra_synonymous‘.‘co_id‘  WHERE ‘sider‘.‘pubchem_id‘=";
try {
Class.forName("com.mysql.jdbc.Driver");
conn = DriverManager.getConnection(url+dbName,userName,password);
state = conn.createStatement();
//-------------------------------------------------------------------------
//-------------------------------------------------------------------------
for(int H=0;H<drug_list1.size();H++){
System.out.println(drug_list1.get(H).toString());
//--------------extraire la liste de terme du drug 1
String ListeTerme1 ;
ListeTerme1=ini.get(drug_list1.get(H).toString(), "List_Terme");
terme_Liste1.clear();
String[] tocken1=ListeTerme1.split(",");
for(int zz=0;zz<tocken1.length;zz++)
{if(terme_Liste1.indexOf(tocken1[zz].toString().trim())<0)
{terme_Liste1.add(tocken1[zz].toString().trim());}
}
String IAF=" SELECT * FROM ‘meddra_information_content‘ WHERE ‘meddra_term_id‘ LIKE ’";
//---extraire la liste de IAF pour la liste 1 des terms  du drug 1
ResultSet  result3,result4;
IAF1.clear();
for(int x=0;x<terme_Liste1.size();x++){
String IAF1_str=IAF+terme_Liste1.get(x)+"’";
result3 = state.executeQuery(IAF1_str);
int zz=-1;
while (result3.next()) {
String tmp=result3.getString(2).trim();
IAF1.add(tmp);
zz=1;
}// while result
if(zz==-1){IAF1.add("0.0");}
}// for x
//---------extraire la liste de terme du drug 2
for(int L=0;L<drug_list2.size();L++){
System.out.println(drug_list2.get(L).toString());

String ListeTerme2 ;
ListeTerme2=ini.get(drug_list2.get(L).toString(), "List_Terme");
terme_Liste2.clear();
String[] tocken2=ListeTerme2.split(",");
for(int z=0;z<tocken2.length;z++){
if(terme_Liste2.indexOf(tocken2[z].toString().trim())<0)
{terme_Liste2.add(tocken2[z].toString().trim());}
}
//----------------
```

```
//-------------------------
//---extraire la liste de IAF pour la liste 2 des terms  du drug 2
IAF2.clear();
for(int x=0;x<terme_Liste2.size();x++){
String IAF1_str=IAF+terme_Liste2.get(x)+"'";
result4 = state.executeQuery(IAF1_str);
int zz=-1;
while (result4.next()) {
String tmp=result4.getString(2).trim();
IAF2.add(tmp);
zz=1;
}// while result
if(zz==-1){IAF2.add("0.0");}
}// for x
//-------------------------------------
//-----------------------------------
// le calcul de la similarité ---------

double total_similarity=0.0, maqam=0.0, kasr1=0.0,kasr2=0.0;
String simi_query="SELECT `similarity` FROM `meddra_similarity`
WHERE `terme1` LIKE '";
@SuppressWarnings("unused")
ResultSet  result5,result6,result7,result52;
///---
for(int x=0;x<terme_Liste1.size();x++){
for(int y=0;y<terme_Liste2.size();y++){
double dot_product=0.0;
/*
*
* connecter à la base de donner pour récuperer
Sim(terme_Liste1.get(x),terme_Liste2.get(y));
*
* */
@SuppressWarnings("unused")
String tmp=simi_query+terme_Liste1.get(x)+"' AND `terme2` LIKE '"
+terme_Liste2.get(y)+"'";
String tmp2=simi_query+terme_Liste2.get(y)+"' AND `terme2` LIKE '"
+terme_Liste1.get(x)+"'";

if(terme_Liste1.get(x)==terme_Liste2.get(y))dot_product=1.0;
else{
String zzz="";
zzz=inisim.get(terme_Liste1.get(x),terme_Liste2.get(y));

if((zzz==null)||(zzz==="")) dot_product=0.0;
else dot_product=Double.valueOf(zzz  );
}
double IAF_product= Double.valueOf(IAF1.get(x).toString())*
```

```
Double.valueOf(IAF2.get(y).toString());
maqam=maqam+ dot_product*IAF_product;

}// for y
}// for x
//-----------
kasr1=0.0;
for(int x=0;x<terme_Liste1.size();x++){
for(int y=0;y<terme_Liste1.size();y++){
double dot_product=0.0;
/*
*
* connecter à la base de donner pour récuperer
Sim(terme_Liste1.get(x),terme_Liste1.get(y));
*
* */
String tmp=simi_query+terme_Liste1.get(x)+"' AND `terme2` LIKE '"
+terme_Liste1.get(y)+"'";

//  System.out.println(zzz);
if(terme_Liste1.get(x)==terme_Liste1.get(y))dot_product=1.0;
else{
String zzz="";
zzz=inisim.get(terme_Liste1.get(x),terme_Liste1.get(y));

if((zzz==null)||(zzz==""))  dot_product=0.0;
else dot_product=Double.valueOf(zzz  );
}
/*String xx="nonmodified";
result6=state.executeQuery(tmp);
while (result6.next()) {
String    zzz=result6.getString(1).trim().replace(',', '.');
xx="modified";
dot_product=Double.valueOf(zzz);
}

if(xx.equals("modified")==false){
tmp=simi_query+terme_Liste1.get(y)+"' AND `terme2` LIKE '"
+terme_Liste1.get(x)+"'";
result6=state.executeQuery(tmp);
xx="nonmodified";
while (result6.next()) {
String    zzz=result6.getString(1).trim().replace(',', '.');
xx="modified"; dot_product=Double.valueOf(zzz);
}// while result

}// while result
*/
```

```
double IAF_product= Double.valueOf(IAF1.get(x).toString())
*Double.valueOf(IAF1.get(y).toString());
kasr1=kasr1+ dot_product*IAF_product;

}// for y
}// for x
kasr1=Math.sqrt(kasr1);

//-----------Kasre2-
kasr2=0.0;
for(int x=0;x<terme_Liste2.size();x++){

for(int y=0;y<terme_Liste2.size();y++){
double dot_product=0.0;
/*
*
* connecter à la base de donner pour récuperer
Sim(terme_Liste2.get(x),terme_Liste2.get(y));
*
* */
String tmp=simi_query+terme_Liste2.get(x)+"' AND `terme2` LIKE '"
+terme_Liste2.get(y)+"'";
if(terme_Liste2.get(x)==terme_Liste2.get(y))dot_product=1.0;
else{
String zzz="";
zzz= inisim.get(terme_Liste2.get(x),terme_Liste2.get(y));
if((zzz==null)||(zzz==""")) dot_product=0.0;
else dot_product=Double.valueOf(zzz  );
}

double IAF_product= Double.valueOf(IAF2.get(x).toString())
*Double.valueOf(IAF2.get(y).toString());
kasr2=kasr2+ dot_product*IAF_product;

}// for y
}// for x
kasr2=Math.sqrt(kasr2);
//--------
total_similarity= maqam/( kasr1*kasr2);
if(total_similarity>=1.0)
{

}
else{
NumberFormat formater3 = new DecimalFormat("#.###");
System.out.println("Semantic Similarity Between Drug ID1: "
+drug_list1.get(H)+" and
Drug ID2: "+drug_list2.get(L)+" = "+formater3.format(total_similarity));
```

```
}
}// H
}// K
state.close();
conn.close();
}
catch (ClassNotFoundException e) {
}
catch (SQLException e) {
}
}
}
```

Index

Bibliographie

[ABB+00] Michael Ashburner, Catherine Ball, Judith Blake, David Botstein, Heather Butler, Michael Cherry, Allan Davis, Kara Dolinski, Selina Dwight, Janan Eppig, Midori Harris, David Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel Richardson, Martin Ringwald, Gerald Rubin, and Gavin Sherlock. Gene ontology : tool for the unification of biology. *Nature Genetics*, 25 :25–29, 2000.

[ACDI06] Musa H. Asyali, Dilek Colak, Omer Demirkaya, and Mehmet S. Inan. Gene expression profile classification : A review. *Current Bioinformatics*, 1 :55–73, 2006.

[AFH08] Snchez L. Garca S. del Jesus M. Ventura S. Garrell J. Otero J. Romero C. Bacardit J. Rivas V. Fernndez J. Alcala-Fdez, J. and F. Herrera. Keel : a software tool to assess evolutionary algorithms for data mining problems. *Soft Comput.*, 13 :307–318, October 2008.

[AKG+91] MD Adams, JM Kelley, JD Gocayne, M Dubnick, MH Polymeropoulos, H Xiao, CR Merril, A Wu, B Olde, RF Moreno, and al. et. Complementary dna sequencing : expressed sequence tags and human genome project. *Science*, 252(5013) :1651–1656, 1991.

[AMI] The AMIGO database. `http://amigo.geneontology.org`.

[AMM44] Oswald T. Avery, Colin M. MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.*, 79(2), 1944.

[AMP+10] Panagiotis Alexiou, Manolis Maragkakis, Giorgio L. Papadopoulos, Victor A. Simmosis, Lin Zhang, and Artemis G. Hatzigeorgiou. The DIANA-mirExTra web server : from gene expression data to microRNA function. *PloS one*, 5(2) :e9171+, February 2010.

[ARL06] Adrian Alexa, Jorg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, 22(13) :1600–1607, 2006.

[AS04] Boris Adryan and Reinhard Schuh. Gene-Ontology-based clustering of gene expression data. *Bioinformatics*, 20(16) :2851–2852, 2004.

[ASAD+07] Fátima Al-Shahrour, Leonardo Arbiza, Hernán Dopazo, Jaime Huerta-Cepas, Pablo Mínguez, David Montaner, and Joaquín Dopazo. From genes to functional classes in the study of biological systems. *BMC bioinformatics*, 8 :114+, April 2007.

[ASDUD04] Fatima Al-Shahrour, Ramon Diaz-Uriarte, and Joaquin Dopazo. FatiGO : a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4) :578–580, 2004.

[BA01] Quackenbush J Sherlock G Spellman P Stoeckert C Aach J Ansorge W Ball CA Causton HC Gaasterland T Glenisson P Holstege FC Kim IF Markowitz V Matese JC Parkinson H Robinson A Sarkans U Schulze-Kremer S Stewart J Taylor R Vilo J Vingron M Brazma A, Hingamp P. Minimum information about a microarray experiment (miame)-toward standards for microarray data. *Nat Genet*, 29(4) :365–371, 2001.

[BAB05] Olivier Bodenreider, Marc Aubry, and Anita Burgun. Non-lexical approaches to identifying associative relations in the gene ontology. In *in the Gene Ontology. PBS 2005*, pages 91–102, 2005.

[BBR+07] S. Blachon, J. Besson, C. Robardet, J-F. Boulicaut, O. Gandrillon, and R.G. Pensa. Clustering formal concepts to discover biologically relevant knowledge from gene expression data. *In silico Biology*, 7(4-5) :467–483, 2007.

[BBST+11] Emmanuel Bresso, Sidahmed Benabderrahmane, Malika Smaïl-Tabbone, Gino Marchetti, Arnaud Sinan Karaboga, Michel Souchet, Amedeo Napoli, and Marie-Dominique Devignes. Use of domain knowledge for dimension reduction : application to mining of drug side effects. In Ana Fred, editor, *KDIR 2011- Third International Conference on Knowledge Discovery and Information Retrieval*, Paris, France, 2011. INSTICC.

[BCG+05] Stephen Blott, Fabrice Camous, Cathal Gurrin, Gareth J. F. Jones, and Alan F. Smeaton. On the use of clustering and the mesh controlled vocabulary to improve medline abstract search. In *CORIA*, pages 41–56, 2005.

[BDH+09] Daniel Barrell, Emily Dimmer, Rachael P. Huntley, David Binns, Claire O'Donovan, and Rolf Apweiler. The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucl. Acids Res.*, 37(suppl1) :D396–403, 2009.

[BdlFM02] Paul Brazhnik, Alberto de la Fuente, and Pedro Mendes. Gene networks : how to put the function in genomics. *Trends in Biotechnology*, 20(11) :467 – 472, 2002.

[BDST+09] Sidahmed Benabderrahmane, Marie-Dominique Devignes, Malika Smaïl-Tabbone, Amedeo Napoli, Olivier Poch, Ngoc-H. Nguyen, and Wolfgang Raffelsberger. Analyse de données transcriptomiques : Modélisation floue de profils dexpression différentielle et analyse fonctionnelle. In *INFORSID*, pages 413–428, 2009.

[BDY99] Amir Ben-Dor and Zohar Yakhini. Clustering gene expression patterns. In *Proceedings of the third annual international conference on Computational molecular biology*, RECOMB '99, pages 33–42, New York, NY, USA, 1999. ACM.

[BE06] Tanya Barrett and Ron Edgar. Gene expression omnibus : microarray data storage, submission, retrieval, and analysis. *Methods in enzymology*, 411 :352–369, 2006.

[Ben06] Younes Bennani. *Apprentissage Connexionniste*. Lavoisier, Paris, 2006.

[Bez81] James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 1981.

[BFR+04] Francois Bertucci, Pascal Finetti, Jacques Rougemont, Emmanuelle Charafe-Jauffret, Valery Nasser, Beatrice Loriod, Jacques Camerlo, Rebecca Tagett, Carole Tarpin, Gilles Houvenaeghel, Catherine Nguyen, Dominique Maraninchi, Jocelyne Jacquemier, Remi Houlgatte, Daniel Birnbaum, and Patrice Viens. Gene expression profiling for molecular characterization of inflammatory breast cancer and prediction of response to chemotherapy. *Cancer Research*, 64(23) :8558–8565, 2004.

[BGG+07] Fairouz Benahmed, Isabelle Gross, Dominique Guenot, Frederic Jehan, Elisabeth Martin, Claire Domon-Dell, Thomas Brabletz, Michele Kedinger, Jean Noel Freund, and Isabelle Duluc. The microenvironment controls cdx2 homeobox gene expression in colorectal cancer cells. *The American Journal of Pathology*, 170(2) :733–744, 2007.

[BGL+00] P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceeding National Academy of Science USA*, 97(1) :262–267, 2000.

[BHK08] Emmanuel Blanchard, Mounira Harzallah, and Pascale Kuntz. A generic framework for comparing semantic similarities on a subsumption hierarchy. In *18th European Conference on Artificial Intelligence (ECAI)*, pages 20–24, 2008.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[BJ10] Jean-François Boulicaut and Baptiste Jeudy. Constraint-based Data Mining. In Lior Maimon, Oded ; Rokach, editor, *Data Mining and Knowledge Discovery Handbook*, pages 339–354. Springer, 2010. 2nd ed., 2010, XX, 1285 p. 40 illus., Hardcover ISBN : 978-0-387-09822-7.

[BL01] Pierre Baldi and Anthony D. Long. A bayesian framework for the analysis of microarray expression data : Regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17 :509–519, 2001.

[Bod09] Olivier Bodenreider. Special issue : Biomedical ontology in action. *Applied Ontology*, 4(1) :1–4, 2009.

[BS04a] Tim Beisbarth and Terence P. Speed. Gostat : find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9) :1464–1465, 2004.

[BS04b] Tim Beissbarth and Terence P. Speed. GOstat : find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9) :1464–1465, 2004.

[BSE+04] F. Bertucci, S. Salas, S. Eysteries, V. Nasser, P. Finetti, C. Ginestier, E. Charafe-Jauffret, B. Loriod, L. Bachelart, J. Montfort, G. Victorero, F. Viret, V. Ollendorff, V. Fert, M. Giovaninni, J. R. Delpero, C. Nguyen, P. Viens, G. Monges, D. Birnbaum, and R. Houlgatte. Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. *Oncogene*, 23(7) :1377–1391, February 2004.

[BSTND] Sidahmed Benabderrahmane, Malika Smail-Tabbone, Amedeo Napoli, and Marie-Dominique Devignes. Ontology-based functional classification of genes : evaluation with reference sets and overlap analysis. *IEEE International Conference Bioinformatics and Biomedicine, the Second Workshop on Integrative Data Analysis in Systems Biology. 12-15 Nov 2011, Atlanta-USA.*

[BSTP+10] Sidahmed Benabderrahmane, Malika Smail-Tabbone, Olivier Poch, Amedeo Napoli, and Marie-Dominique Devignes. Intelligo : a new vector-based semantic similarity measure including annotation origin. *BMC Bioinformatics*, 11(1) :588, 2010.

[BSTP+ce] Sidahmed Benabderrahmane, Malika Smail-Tabbone, Olivier Poch, Amedeo Napoli, Ngoc Hoan, Raffelsberger Wolfgang, eric Guerin, Dominique Guenot, and Marie-Dominique Devignes. Benchmarking a new semantic similarity measure using fuzzy clustering and reference sets : Application to cancer expression data. In *11 eme Conference Internationale Francophone sur l'Extraction et la Gestion des Connaissances EGC 2011. Atelier Fouille de données complexes.*, 25-28 janvier 2011, Brest, France.

[Bus00] SA Bustin. Absolute quantification of mrna using real-time reverse transcription polymerase chain reaction assays. *Journal of Molecular Endocrinology*, 25(2) :169–193, 2000.

[BW07] Markus Brameier and Carsten Wiuf. Co-clustering and visualization of gene expression data and gene ontology terms for saccharomyces cerevisiae using self-organizing maps. *J. of Biomedical Informatics*, 40(2) :160–173, 2007.

[BWG+04] Elizabeth I. Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J. Michael Cherry, and Gavin Sherlock. Go : :termfinder-open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18) :3710–3715, 2004.

[BWKR03] Gabriel F. Berriz, James V. White, Oliver D. King, and Frederick P. Roth. Gofish finds genes with combinations of gene ontology attributes. *Bioinformatics*, 19(6) :788–789, 2003.

[BWPS78] G N Buell, M P Wickens, F Payvar, and R T Schimke. Synthesis of full length cdnas from four partially purified oviduct mrnas. *The Journal of Biological Chemistry*, 253(7) :2471–2482, 1978.

[BY01] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29 :1165–1188, 2001.

[Cam86] James R Campbell. An ambulatory information system serving the needs of clinical practice : Costar v. In *Proc Annu Symp Comput Appl Med Care.*, pages 141–146, 1986.

[CCT10] S. S. Choi, S. H. Cha, and C. Tappert. A Survey of Binary Similarity and Distance Measures. *Journal on Systemics, Cybernetics and Informatics*, 8(1) :43–48, 2010.

[ces] Collaborative Evaluation of GO-based Semantic Similarity Measures. `http://xldb.di.fc.ul.pt/tools/cessm/`.

[CFC+09] Paulo Carvalho, Juliana Fischer, Emily Chen, Gilberto Domont, Maria Carvalho, Wim Degrave, John Yates, and Valmir Barbosa. Go explorer : A gene-ontology tool to aid in the interpretation of shotgun proteomics data. *Proteome Science*, 7(1) :6, 2009.

[CG08] Nicole Cloonan and Sean Grimmond. Transcriptome content and dynamics at single-nucleotide resolution. *Genome Biology*, 9(9) :234, 2008.

[Cha07] Sung Hyuk Cha. Comprehensive survey on distance similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4) :300–307, 2007.

[CKG+08] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science*, 321(5886) :263–266, 2008.

[Cla99] Jean-Michel Claverie. Computational methods for the identification of differential and coordinated gene expression. 8(10) :1821–1832, 1999.

[CMB07] Julie Chabalier, Jean Mosser, and Anita Burgun. A transversal approach to predict gene product networks from ontology-based similarity. *BMC Bioinformatics*, 8(1) :235, 2007.

[Coh60] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1) :37–46, 1960.

[Con10] The Gene Ontology Consortium. The Gene Ontology in 2010 : extensions and refinements. *Nucl. Acids Res.*, 38(suppl-1) :D331–335, 2010.

[CSC07] Francisco M. Couto, Mario J. Silva, and Pedro M. Coutinho. Measuring semantic similarity between gene ontology terms. *Data Knowl. Eng.*, 61(1) :137–152, 2007.

[CSTB+08] Adrien Coulet, Malika Smail-Tabbone, Pascale Benlian, Amedeo Napoli, and Marie-Dominique Devignes. Ontology-guided data preparation for discovering genotype-phenotype relationships. *BMC Bioinformatics*, 9(Suppl 4) :S3, 2008.

[DBD+04] E. Demir, O. Babur, U. Dogrusoz, A. Gursoy, A. Ayaz, G. Gulesir, G. Nisanci, and R. Cetin-Atalay. An ontology for collaborative construction and analysis of cellular pathways. *Bioinformatics*, 20(3) :349–356, 2004.

[DBG] The DBGET database retrieval system. `http://www.genome.jp/dbget/`.

[DBW04] Jennifer G. Dy, Carla E. Brodley, and Stefan Wrobel. Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5 :845–889, 2004.

[DDCG09] Alpha Diallo, Ahlame Douzal Chouakria, and Francoise Giroud. Which distance for the identification and the differentiation of cell cycle expressed genes. *Advances in Intelligent Data Analysis VIII*, 5772 :273–284, 2009.

[DDG08] Alpha Diallo, Ahlame Douzal, and Françoise Giroud. Classification adaptative de séries temporelles : application à l'identification des gènes exprimés au cours du cycle cellulaire. In *EGC*, pages 487–498, 2008.

[DFS02] Sandrine Dudoit, Jane Fridlyand, and Terence P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the american statistical association*, 97(457) :77–87, 2002.

[DK03] Doulaye Dembélé and Philippe Kastner. Fuzzy c-means method for clustering microarray data. *Bioinformatics*, 19(8) :973–980, 2003.

[DLS00] Patrik Dhaeseleer, Shoudan Liang, and Roland Somogyi. Genetic network inference : from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8) :707–726, 2000.

[dMAD+10] Paula de Matos, Rafael Alcántara, Adriano Dekker, Marcus Ennis, Janna Hastings, Kenneth Haug, Inmaculada Spiteri, Steve Turner, and Christoph Steinbeck. Chemical Entities of Biological Interest : an update. *Nucleic Acids Research*, 38(suppl 1) :D249–D254, January 2010.

[Dra03] Sorin Draghici. *Data Analysis Tools for DNA Microarrays*. Chapman & Hall/CRC, June 2003.

[DSH+03] Glynn Dennis, Brad Sherman, Douglas Hosack, Jun Yang, Wei Gao, H Lane, and Richard Lempicki. David : Database for annotation, visualization, and integrated discovery. *Genome Biology*, 4(9) :R60, 2003. A previous version of this manuscript was made available before peer review at http ://genomebiology.com/2003/4/5/P3.

[ea06] David Allis et al. *CSHL EPIGENETICS*. Cold Spring Harbour Laboratory Press, CSHL, NY, USA, 2006.

[ENB+07] Mohamed Elati, P. Neuvial, M. Bolotin, E. Barillot, C. Rouveirol, and F. Radvany. Licorn : learning co-operative regulation networks from expression data. *Bioinformatics*, 23 :2497–2414, 2007.

[ENS+09] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. Gorilla : a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics*, 10(1) :48, 2009.

[ESBB98] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25) :14863–14868, December 1998.

[evi] The Gene Ontology Evidence Tree. `http://www.geneontology.org/GO.evidence.tree.shtml`.

[FL08] Ana L. N. Fred and José M. N. Leito. A new cluster isolation criterion based on dissimilarity increments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25 :944–958, 2008.

[Flo04] Luciano Floridi. Outiline of a theory of strongly semantic information. *Minds Mach.*, 2004.

[FMSB+] Robert D. Finn, Jaina Mistry, Benjamin Schuster-Bockler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, Mhairi Marshall, Ajay Khanna, Richard Durbin, Sean R. Eddy, Erik L. L. Sonnhammer, and Alex Bateman. Pfam : clans, web tools and services. *Nucl. Acids Res.*, 34.

[FSRL01] Helene Fischer, Roger Stenling, Carlos Rubio, and Annika Lindblom. Differential expression of aquaporin 8 in human colonic epithelial cells and colorectal tumors. *BMC Physiology*, 1(1) :1, 2001.

[GAM+03] P. Glenisson, P. Antal, J. Mathys, Y. Moreau, and B. De Moor. Evaluation of the vector space representation in text-based gene clustering. In *Proc of the Eighth Ann Pac Symp Biocomp (PSB 2003*, pages 391–402, 2003.

[GBee06] S. G. Gregory, K. F. Barlow, and K. E. McLay et el. The dna sequence and biological annotation of human chromosome 1. *Nature*, 441 :315–321, 2006.

[GBRV07] Steffen Grossmann, Sebastian Bauer, Peter N. Robinson, and Martin Vingron. Improved detection of overrepresentation of gene ontology annotations with parent child analysis. 23(22) :3024–3031, 2007.

[GE02] Audrey Gasch and Michael Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3(11) :research0059.1–research0059.22, 2002.

[GGMW03] Prasanna Ganesan, Hector Garcia-Molina, and Jennifer Widom. Exploiting hierarchical domain structure to compute similarity. *ACM Trans. Inf. Syst.*, 21(1) :64–93, 2003.

[GL05] M. Grabisch and C. Labreuche. Bi-capacities-ii the choquet integral. *Fuzzy Sets and Systems*, 151(2) :237–259, April 2005.

[GLS+06] Xiang Guo, Rongxiang Liu, Craig D. Shriver, Hai Hu, and Michael N. Liebman. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, 22(8) :967–973, 2006.

[GOv] The Bioconductor GOstats package. `http://bioconductor.org/packages/2.5/bioc/vignettes/GOstats/inst/doc/GOvis.pdf`.

[GSU08] Raffaele Giancarlo, Davide Scaturro, and Filippo Utro. Computational cluster validation for microarray data analysis : experimental assessment of Clest, Consensus Clustering, Figure of Merit, Gap Statistics and Model Explorer. *BMC bioinformatics*, 9(1) :462+, 2008.

[GUE05] Emilie GUERIN. *Intģration des données pour l'analyse de transcriptome : Mise en oeuvre par l'entrepôt GEDAW : Gene Expression DATA WareHouse*. Doctorat, Université Rennes, 2005.

[Guy03] Isabelle Guyon. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3 :1157–1182, 2003.

[GW99] B. Ganter and R. Wille. *Formal Concept Analysis : Mathematical foundations*. Springer, Ganter99, 1999.

[Hat79] Michael A W Hattwick. Computer stored ambulatory record systems in real life practice. In *Proc Annu Symp Comput Appl Med Care.*, pages 761–764, 1979.

[HDS+03a] Douglas Hosack, Glynn Dennis, Brad Sherman, H Lane, and Richard Lempicki. Identifying biological themes within lists of genes with ease. *Genome Biology*, 4(10)R70 :4, 2003.

[HDS+03b] Douglas A. Hosack, Glynn Dennis, Brad T. Sherman, H. Clifford Lane, and Richard A. Lempicki. Identifying biological themes within lists of genes with EASE. *Genome biology*, 4(10) :R70+, 2003.

[Hel02] Michael J. Heller. Dna microarray technology : Devices, systems, and applications. *Annual Review of Biomedical Engineering*, 4(1) :129–153, 2002.

[HP06] Desheng Huang and Wei Pan. Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. *Bioinformatics*, 22(10) :1259–1268, 2006.

[HSL09] Da Wei a. . W. Huang, Brad T. Sherman, and Richard A. Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1) :44–57, December 2009.

[HST+07] Da Huang, Brad Sherman, Qina Tan, Jack Collins, W Gregory Alvord, Jean Roayaei, Robert Stephens, Michael Baseler, H Clifford Lane, and Richard Lempicki. The david gene functional classification tool : a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, 8(9) :R183, 2007.

[IA05] Sandrine Imbeaud and Charles Auffray. The 39 steps in gene expression profiling : critical issues and proposed best practices for microarray experiments. *Drug Discovery Today*, 10(17) :1175 – 1182, 2005.

[IK03] Tamura T Gojobori T Tateno Y Ikeo K, Ishi-i J. Cibex : center for information biology gene expression database. *C R Biol*, 326(10-11) :1079–82, 2003.

[JB10] Lars J. Jensen and Peer Bork. Ontologies in quantitative biology : A basis for comparison, integration, and discovery. *PLoS Biol*, 8(5) :e1000374, 05 2010.

[JC97] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*, pages 9008+, September 1997.

[JEB+97] T. C. James, S. C. Elgin, Mol Cell Biol, A. Schmitz, K. H. Gartemann, J. Fiedler, R. Eichenlaub, V. Sharma, K. Suvarna, R. Meganathan, We Thank H. Skaletsky, F. Lewitter For Help, P. Bain, A. Bortvin, A. De La Chapelle, G. Fink, T. Kawaguchi, H. Lodish, D. Menke, U. Rajbh, R. Reijo, A. Schwartz, C. Sun, C. Tilford For Comments, Joseph L. Derisi, Vishwanath R. Iyer, and Patrick O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338) :680–686, 1997.

[JFB04] Olivier Gandrillon Jean Francois Boulicaut. *Informatique pour l'analyse du transcriptome.* Lavoisier-Hermes, Paris, 1 edition, 2004.

[JKP94] George H. John, Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In *MACHINE LEARNING : PROCEEDINGS OF THE ELEVENTH INTERNATIONAL*, pages 121–129. Morgan Kaufmann, 1994.

[Joh73] Donald B. Johnson. A note on dijkstra's shortest path algorithm. *J. ACM*, 20 :385–388, July 1973.

[JTL+08] Xia Jiang, Jing Tan, Jingsong Li, Saul Kivime, Xiaojing Yang, Li Zhuang, Puay Leng Lee, Mark T.W. Chan, Lawrence W. Stanton, Edison T. Liu, Benjamin N.R. Cheyette, and Qiang Yu. Dact3 is an epigenetic regulator of wnt b catenin signaling in colorectal cancer and is a therapeutic target of histone modifications. *Cancer Cell*, 13(6) :529–541, 2008.

[KCL+10] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 6(343) :343, 2010.

[KD05] Purvesh Khatri and Sorin Draghici. Ontological analysis of gene expression data : current tools, limitations, and open problems. *Bioinformatics*, 21(18) :3587–3595, 2005.

[KDKN09] Mehdi Kaytoue, Sébastien Duplessis, Sergei O. Kuznetsov, and Amedeo Napoli. Two fca-based methods for mining gene expression data. In *Proceedings of the 7th International Conference on Formal Concept Analysis*, ICFCA '09, pages 251–266, Berlin, Heidelberg, 2009. Springer-Verlag.

[KDN08] M Kaytoue, Sébastien Duplessis, and Amedeo Napoli. Using formal concept analysis for extraction of groups of co-expressed genes. *Modeling, Computation and Optimization in Information Systems and Management Sciences (MCO 2008).*, 14 :439–449, 2008.

[KDOK02] Purvesh Khatri, Sorin Draghici, G. Charles Ostermeier, and Stephen A. Krawetz. Profiling gene expression using onto-express. *Genomics*, 79(2) :266 – 270, 2002.

[KEG] The KEGG Pathways database. `http://www.genome.jp/kegg/pathway.html`.

[KFD+03] Oliver D. King, Rebecca E. Foulger, Selina S. Dwight, James V. White, and Frederick P. Roth. Predicting gene function from patterns of annotation. *Genome Research*, 13(5) :896–904, 2003.

[KGF+10] Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, 38(Database issue) :D355–360, January 2010.

[KGS96] Lawrence A. Kelley, Stephen P. Gardner, and Michael J. Sutcliffe. An automated approach for clustering an ensemble of nmr-derived protein structures into conformationally related subfamilies. *Protein Engineering*, 9(11) :1063–1065, 1996.

[KHKL96] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen. SOM PAK : The Self-Organizing Map program package. 1996.

[KJ97] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artif. Intell.*, 97 :273–324, December 1997.

[KKR+04] Hyuk-Chan Kwon, Sung-Hyun Kim, Mee-Sook Roh, Jae-Seok Kim, Hyung-Sik Lee, Hong-Jo Choi, Jin-Sook Jeong, Hyo-Jin Kim, and Tae-Ho Hwang. Gene expression profiling in lymph node-positive and lymph node-negative colorectal cancer. *Diseases of the Colon Rectum*, 47 :141–152, 2004. 10.1007/s10350-003-0032-7.

[KLJ+11] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, Yannick Djoumbou, Roman Eisner, Anchi Guo, and David S. Wishart. Drugbank 3.0 : a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research*, 39(Database-Issue) :1035–1041, 2011.

[KLL04] Dae-Won Kim, Kwang H. Lee, and Doheon Lee. On cluster validity index for estimation of the optimal number of fuzzy clusters. *Pattern Recognition*, 37(10) :2009 – 2025, 2004.

[KMC00] M. Kathleen Kerr, Mitchell Martin, and Gary A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7 :819–837, 2000.

[Kni06] Carole Knibbe. *Structuration des génomes par sélection indirecte de la variabilité mutationnelle : une approche de modélisation et de simulation.* PhD thesis, INSA-Lyon (Institut National des Sciences Appliquées de Lyon), 2006.

[Kos99] B Kosko. Neural networks and fuzzy systems. *Englewood Cliffs, NJ : Prentice-Hall*, 1999.

[KR05] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data : An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics).* Wiley-Interscience, March 2005.

[KS96] Daphne Koller and Mehran Sahami. Toward optimal feature selection. Technical Report 1996-77, Stanford InfoLab, February 1996. Previous number = SIDL-WP-1996-0032.

[KSM00] Yongseog Kim, W. Nick Street, and Filippo Menczer. Feature selection in unsupervised learning via evolutionary search. In *In Proceedings of the Sixth ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 365–369. ACM Press, 2000.

[KWR+01] Javed Khan, Jun S. Wei, Markus Ringner, Lao H. Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R. Antonescu, Carsten Peterson, and Paul S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*, 7(6) :673–679, 2001.

[Kyr08] Antonia Kyriakopoulou. Text classification aided by clustering : a literature review. *Tools in Artificial Intelligence*, pages 233–252, 2008.

[LAJS05] Lovisa Lovmar, Annika Ahlford, Mats Jonsson, and Ann-Christine Syvanen. Silhouette scores for assessment of snp genotype clusters. *BMC Genomics*, 6(1) :35, 2005.

[LBPS] Antonio Di Leva, Roberto Berchi, Gianpiero Pescarmona, and Michele Sonnessa. Analysis and prototyping of biological systems : the abstract biological process model. *International Journal of Information and Technology*, 3(Database-Issue) :216–224.

[LC03] Yuk Fai Leung and Duccio Cavalieri. Fundamentals of cdna microarray data analysis. *Trends Genet*, 19 :649–659, 2003.

[LCH+04] Kevin K. Lin, Darya Chudova, G. Wesley Hatfield, Padhraic Smyth, and Bogi Andersen. Identification of hair cycle-associated genes from time-course gene expression profile data by using replicate variance. *Proceedings of the National Academy of Sciences of the United States of America*, 101(45) :15955–15960, 2004.

[Leb03] Mustapha Lebbah. *Carte topologique pour donn´ees qualitatives : application ‘a la reconnaissance automatique de la densit ´e du trafic routier*. Doctorat, UNIVERSITE de Versailles Saint Quentin-en-Yvelines, 2003.

[LGKFO09] David Landsman, Robert Gentleman, Janet Kelso, and B. F. Francis Ouellette. Database : A new forum for biological databases and curation. *Database*, 09, 2009.

[LGM+05] Jun Lu, Gad Getz, Eric A. Miska, Ezequiel Alvarez-Saavedra, Justin Lamb, David Peck, Alejandro Sweet-Cordero, Benjamin L. Ebert, Raymond H. Mak, Adolfo A. Ferrando, James R. Downing, Tyler Jacks, H. Robert Horvitz, and Todd R. Golub. MicroRNA expression profiles classify human cancers. *Nature*, 435(7043) :834–838, June 2005.

[LH03] Ying Lu and Jiawei Han. Cancer classification using gene expression data. *Inf. Syst.*, 28 :243–268, June 2003.

[LHM93] D. Lindberg, B. Humphreys, and A. McCray. The unified medical language system. *Methods of Information in Medicine*, 32(4) :281–291, 1993.

[LHS+04] Homin K. Lee, Amy K. Hsu, Jon Sajdak, Jie Qin, and Paul Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome Research*, 14(6) :1085–1094, 2004.

[LHW05] Hongfang Liu, Zhang Zhi Hu, and Cathy Wu. Dyngo a tool for visualizing and mining of gene ontology and its associations. *BMC Bioinformatics*, 6(1) :201, 2005.

[Lin98] Dekang Lin. An information-theoretic definition of similarity. In *ICML '98 : Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[Lip] Carolyn Lipscomb. Medical Subject Headings MeSH. `http://www.nlm.nih.gov/mesh/`.

[LKF+04] Nada Lavrac, Branko Kavsek, Peter Flach, Ljupco Todorovski, and Stefan Wrobel. Subgroup discovery with cn2-sd. *Journal of Machine Learning Research*, 5 :153–188, 2004.

[LKS05] Jason Lee, Gurpreet Katari, and Ravi Sachidanandam. Gobar : A gene ontology based analysis and visualization tool for gene sets. *BMC Bioinformatics*, 6(1) :189, 2005.

[LLZ] Yin Liu, Nianjun Liu, and Hongyu Zhao. Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, 21(15) :3279–3285.

[LMLB+04] Nolwenn Le Meur, Guillaume Lamirault, Audrey Bihouee, Marja Steenman, Helene Bedrine-Ferran, Raluca Teusan, Gerard Ramstein, and Jean J. Leger. A dynamic, web-accessible resource to process raw microarray scan data into consolidated gene expression values : importance of replication. *Nucleic Acids Research*, 32(18) :5349–5358, 2004.

[LRB09a] M-J. Lesot, M. Rifqi, and H. Benhadda. Similarity measures for binary and numerical data : a survey. *Int. J. Knowl. Eng. Soft Data Paradigm.*, 1 :63–84, December 2009.

[LRB09b] MJ. Lesot, M. Rifqi, and H. Benhadda. Similarity measures for binary and numerical data : a survey. *Int. J. Knowl. Eng. Soft Data Paradigm.*, 1 :63–84, December 2009.

[LSBG03] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the Gene Ontology : the relationship between sequence and annotation. *Bioinformatics*, 19(10) :1275–1283, 2003.

[LSSM08] Wei-Nchich Lee, Nigam Shah, Karanjot Sundlass, and Mark Musen. Comparison of ontology-based semantic-similarity measures. *AMIA Annu Symp Proceedings*, V2008 :384–388, 2008.

[LY00] Tong Ihn Lee and Richard A. Young. Transcription of eukaryotic protein-coding genes. *Annual Review of Genetics*, 34(1) :77–137, 2000.

[Mac67] J. B. Macqueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[MBR+04] David Martin, Christine Brun, Elisabeth Remy, Pierre Mouren, Denis Thieffry, and Bernard Jacq. Gotoolbox : functional analysis of gene datasets based on gene ontology. *Genome Biology*, 5(12) :R101, 2004.

[McK07] V. McKusick. Mendelian Inheritance in Man and Its Online Version, OMIM. *The American Journal of Human Genetics*, 80(4) :588–604, April 2007.

[MDNST08] Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, and Malika Smail-Tabbone. Many-valued concept lattices for conceptual clustering and information retrieval. In *Proceeding of the 2008 conference on ECAI 2008 : 18th European*

Conference on Artificial Intelligence, pages 127–131, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.

[Mer08] Gary H. Merrill. The MedDRA paradox. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 470–474, 2008.

[Meu05] Nolwenn Le Meur. *De l'Acquisition des Données de Puces à ADN vers leur Interprétation : Importance du Traitement des Données Primaires.* Doctorat, Université De Nantes, 2005.

[Mic08] Magali Michaut. *Analyse de données transcriptome et protéome pour l'étude des réponses aux stress oxydants et aux métaux lourds.* Doctorat, Université PARIS 11, 2008.

[Mil95] George A. Miller. Wordnet : A lexical database for english. *Communications of the ACM*, 38 :39–41, 1995.

[MIM+10] Norikatsu Miyoshi, Hideshi Ishii, Koshi Mimori, Fumiaki Tanaka, Kennichi Nagai, Mamoru Uemura, Mitsugu Sekimoto, Yuichiro Doki, and Masaki Mori. Atp11a is a novel predictive marker for metachronous metastasis of colorectal cancer. *Oncology Reports*, 23(2) :505–510, 2010.

[MMS+05] Michael Maurer, Robert Molidor, Alexander Sturn, Juergen Hartler, Hubert Hackl, Gernot Stocker, Andreas Prokesch, Marcel Scheideler, and Zlatko Trajanoski. Mars : Microarray analysis, retrieval, and storage system. *BMC Bioinformatics*, 6(1) :101, 2005.

[MNS10] Brigham H. Mecham, Peter S. Nelson, and John D. Storey. Supervised normalization of microarrays. *Bioinformatics*, 26(10) :1308–1315, 2010.

[MO04] Sara C. Madeira and Arlindo L. Oliveira. Biclustering algorithms for biological data analysis : a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1 :24–45, 2004.

[NAM08a] A. Nagar and H. Al-Mubaid. Using path length measure for gene clustering based on similarity of annotation terms. In *Computers and Communications, 2008. ISCC 2008. IEEE Symposium on*, pages 637–642, July 2008.

[NAM08b] Anurag Nagar and Hisham Al-Mubaid. A new path length measure based on go for gene similarity with evaluation using sgd pathways. In *Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems (CBMS 08)*, pages 590–595, Washington, DC, USA, 2008. IEEE Computer Society.

[Nap05] A Napoli. A smooth introduction to symbolic methods for knowledge discovery. *Handbook of Categorization in Cognitive Science.*, pages 913–933, 2005.

[NCB] The NCBI gene2go file. `ftp://ftp.ncbi.nih.gov/gene/DATA/gene2go.gz`.

[ODI08] Razib M. Othman, Safaai Deris, and Rosli M. Illias. A genetic similarity algorithm for searching the gene ontology terms and annotating anonymous protein sequences. *J. of Biomedical Informatics*, 41(1) :65–81, 2008.

[OLH08] Kristian Ovaska, Marko Laakso, and Sampsa Hautaniemi. Fast gene ontology based clustering for microarray experiments. *BioData Mining*, 1(1) :11, 2008.

[Ova] The csbl.go package. `http://csbi.ltdk.helsinki.fi/anduril/`.

[P08] Genome Information Integration Project and H-Invitational 2. The h-invitational database (h-invdb), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Research*, 36(suppl 1) :D793–D799, 2008.

[Pav03] Paul Pavlidis. Using anova for gene selection from microarray studies of the nervous system. *Methods*, 31(4) :282 – 289, 2003. Candidate Genes from DNA Array Screens : application to neuroscience.

[PB04] R. Pensa and J.F. Boulicaut. Boolean preperty encoding for local set pattern discovery : an application to gene expressiondata analysis, local pattern detection. *Springer Verlag*, pages 115–134, 2004.

[PCGP06] Alexander Ploner, Stefano Calza, Arief Gusnanto, and Yudi Pawitan. Multidimensional local false discovery rate for microarray studies. *Bioinformatics*, 22(5) :556–565, 2006.

[Pev09] Jonathan Pevsner. *Bioinformatics and functional genomics*. Wiley-Blackwell, 2009.

[pfa] The Pfam_C October 2009 release file. `ftp://ftp.sanger.ac.uk/pub/databases/Pfam/releases/Pfam24.0/Pfam-C.gz`.

[PFB+08] Catia Pesquita, Daniel Faria, Hugo Bastos, Antonio Ferreira, Andre Falcao, and Francisco Couto. Metrics for go based protein semantic similarity : a systematic evaluation. *BMC Bioinformatics*, 9(Suppl 5) :S4, April 2008.

[PFF+09] Catia Pesquita, Daniel Faria, Andre O. Falcao, Phillip Lord, and Francisco M. Couto. Semantic similarity in biomedical ontologies. *PLoS Comput Biol*, 5(7) :e1000443, 07 2009.

[PHW+07] Serguei S. Pakhomov, Harry Hemingway, Susan A. Weston, Steven J. Jacobsen, Richard Rodeheffer, and Véronique L. Roger. Epidemiology of angina pectoris : role of natural language processing of the medical record. *American heart journal*, 153(4) :666–673, April 2007.

[PJS+10] R. M. Parry, W. Jones, T. H. Stokes, J. H. Phan, R. A. Moffitt, H. Fang, L. Shi, A. Oberthuer, M. Fischer, W. Tong, and M. D. Wang. k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *The Pharmacogenomics Journal*, 10(4) :292–309, August 2010.

[PKM06] Mihail Popescu, James M. Keller, and Joyce A. Mitchell. Fuzzy measures on the gene ontology for gene product similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(3) :263–274, 2006.

[Pol04] Nicola Polettini. The vector space model in information retrieval- term weighting problem. *Entropy*, pages 1–9, 2004.

[PPCP07] Elena Perelman, Alexander Ploner, Stefano Calza, and Yudi Pawitan. Detecting differential expression in microarray data : comparison of optimal procedures. *BMC Bioinformatics*, 8(1) :28, 2007.

[PPFC09] Catia Pesquita, Delphine Pessoa, Daniel Faria, and Francisco Couto. Cessm : Collaborative evaluation of semantic similarity measures. In *JB2009 : Challenges in Bioinformatics*, November 2009.

[Qua01] John Quackenbush. Computational analysis of microarray data . *Nature Reviews Genetics*, 2(6) :418–427, June 2001.

[RAM+08] Nitzan Rosenfeld, Ranit Aharonov, Eti Meiri, Shai Rosenwald, Yael Spector, Merav Zepeniuk, Hila Benjamin, Norberto Shabes, Sarit Tabak, Asaf Levy, and et al. Micrornas accurately identify cancer tissue origin. *Nature Biotechnology*, 26(4) :462–469, 2008.

[RBH09] Mark F. Rogers and Asa Ben-Hur. The use of gene ontology evidence codes in preventing classifier assessment bias. *Bioinformatics*, 25(9) :1173–1177, 2009.

[Res95] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453, 1995.

[RGF] Bastien Rance, Jean-François Gibrat, and Christine Froidevaux. An adaptive combination of matchers : Application to the mapping of biological ontologies for genome annotation. In Norman Paton, Paolo Missier, and Cornelia Hedeler, editors, *Data Integration in the Life Sciences*, volume 5647 of *Lecture Notes in Computer Science*, pages 113–126. Springer Berlin Heidelberg.

[RHNV07] Mohamed Hacene Rouane, Marianne Huchard, Amedeo Napoli, and Petko Valtchev. A proposal for combining formal concept analysis and description logics for mining relational data. *ICFCA 2007*, pages 51–65, 2007.

[RJGS05] Merja Ruutu, Bo Johansson, Reidar Grenman, and Stina Syrjaumlnen. Two different global gene expression profiles in cancer cell lines established from etiologically different oral carcinomas. *Oncol Rep*, 14(6) :1511–1157, 2005.

[RM03] Cornelius Rosse and José L. V. Mejino. A reference ontology for biomedical informatics : the foundational model of anatomy. *Journal of Biomedical Informatics*, 36(6) :478 – 500, 2003. Unified Medical Language System.

[RMBB89] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(1) :17–30, Jan/Feb 1989.

[Rou87] Peter Rousseeuw. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1) :53–65, 1987.

[RSA00] Soumya Raychaudhuri, Joshua M. Stuart, and Russ B. Altman. Principal components analysis to summarize microarray experiments : Application to sporulation time series. In *in Pacific Symposium on Biocomputing*, pages 452–463, 2000.

[RWDD08] Seung Yon Rhee, Valerie Wood, Kara Dolinski, and Sorin Draghici. Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, 9(7) :509–515, 2008.

[SAT+05] Alexander Statnikov, Constantin F. Aliferis, Ioannis Tsamardinos, Douglas Hardin, and Shawn Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5) :631–643, 2005.

[SBSS11] Fran Supek, Matko Bosnjak, Nives Skunca, and Tomislav Smuc. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, 6(7) :e21800, 07 2011.

[SBVdFdP+07] Jacob Sabates-Bellver, Laurens G. Van der Flier, Mariagrazia de Palo, Elisa Cattaneo, Caroline Maake, Hubert Rehrauer, Endre Laczko, Michal A. Kurowski, Janusz M. Bujnicki, Mirco Menigatti, Judith Luz, Teresa V. Ranalli, Vito Gomes, Alfredo Pastorelli, Roberto Faggiani, Marcello Anti, Josef Jiricny, Hans Clevers, and Giancarlo Marra. Transcriptome profile of human colorectal adenomas. *Molecular Cancer Research*, 5(12) :1263–1275, 2007.

[Sch10] Andreas Schlicker. *Ontology-based Similarity Measures and their Application in Bioinformatics*. Thèse de doctorat, UNIVERSITÄT DES SAARLANDES, 2010.

[SDRL06] Andreas Schlicker, Francisco Domingues, Jorg Rahnenfuhrer, and Thomas Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, 7(1) :302, 2006.

[SFSZ05] Nora Speer, Holger Frohlich, Christian Spieth, and Andreas Zell. Functional grouping of genes using spectral clustering and gene ontology. In *In Proceedings of the IEEE International Joint Conference on Neural Networks*, pages 298–303. IEEE Press, 2005.

[SHH+06] Rachel Sealfon, Matthew Hibbs, Curtis Huttenhower, Chad Myers, and Olga Troyanskaya. Golem : an interactive graph-based gene-ontology navigation and analysis tool. *BMC Bioinformatics*, 7(1) :443, 2006.

[SHT+07] Brad Sherman, Da Huang, Qina Tan, Yongjian Guo, Stephan Bour, David Liu, Robert Stephens, Michael Baseler, H Clifford Lane, and Richard Lempicki. David knowledgebase : a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics*, 8(1) :426, 2007.

[SLA10] Andreas Schlicker, Thomas Lengauer, and Mario Albrecht. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. *Bioinformatics*, 26(18) :i561–i567, 2010.

[Slo02] Donna K. Slonim. From patterns to pathways : gene expression data analysis comes of age. *Nature genetics*, 32 Suppl :502–508, December 2002.

[SM58] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28 :1409–1438, 1958.

[SM83] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[SNK07] Laszlo Szathmary, Amedeo Napoli, and Sergei O. Kuznetsov. Zart : A multifunctional itemset mining algorithm. In Peter W. Eklund, Jean Diatta, and Michel Liquiere, editors, *CLA*, volume 331 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2007.

[SS04] Gordon K. Smyth and Gordon K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol*, 3, 2004.

[SSP+05] Jose L. Sevilla, Victor Segura, Adam Podhorski, Elizabeth Guruceaga, Jose M. Mato, Luis A. Martinez-Cruz, Fernando J. Corrales, and Angel Rubio. Correlation between gene expression and go semantic similarity. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 2(4) :330–338, 2005.

[SSZ04] Nora Speer, Christian Spieth, and Andreas Zell. A memetic co-clustering algorithm for gene expression profiles and biological annotation. *Proceedings of the 2004 Congress on Evolutionary Computation (CEC 2004), Portland, USA*, 2 :1631–1638, 2004.

[ST00] Noam Slonim and Naftali Tishby. Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 208–215, New York, NY, USA, 2000. ACM.

[STM+05] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43) :15545–15550, 2005.

[Sto05] Roland B. Stoughton. Application of dna microarrays in biology. *Annual Review of Biochemistry*, 74 :53–82, 2005.

[SWS98] Claude E. Shannon, Warren Weaver, and Shannon. *The Mathematical Theory of Communication*. University of Illinois Press, September 1998.

[THC+99] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22 :281 – 285, 1999.

[THNC02] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10) :6567–6572, 2002.

[Tib06] Robert Tibshirani. A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics*, 7(1) :106, 2006.

[TK06] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition, Third Edition*. Academic Press, Inc., Orlando, FL, USA, 2006.

[TN06] Angela Torres and Juan Nieto. Fuzzy logic in medicine and bioinformatics. *Journal of Biomedicine and Biotechnology*, 2006 :1 – 7, 2006.

[Tol05] Marggie D. Gonzalez Toledo. *A COMPARISON IN CLUSTER VALIDATION TECHNIQUES*. Doctorat, UNIVERSITY OF PUERTO RICO, 2005.

[TOR+01a] G. C. Tseng, M. K. Oh, L. Rohlin, J. C. Liao, and W. H. Wong. Issues in cDNA microarray analysis : quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res*, 29(12) :2549–2557, 2001.

[TOR+01b] George C. Tseng, Min-Kyu Oh, Lars Rohlin, James C. Liao, and Wing Hung Wong. Issues in cdna microarray analysis : quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*, 29(12) :2549–2557, 2001.

[TOTZ01] Jeffrey G. Thomas, James M. Olson, Stephen J. Tapscott, and Lue Ping Zhao. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research*, 11(7) :1227–1236, 2001.

[TP09] George Tsatsaronis and Vicky Panagiotopoulou. A generalized vector space model for text retrieval based on semantic relatedness. In *EACL '09 : Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics : Student Research Workshop*, pages 70–78, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[TSM+99] Pablo Tamayo, Donna Slonim, Jill Mesirov, Qing Zhu, Sutisak Kitareewan, Ethan Dmitrovsky, Eric S. Lander, and Todd R. Golub. Interpreting patterns of gene expression with self-organizing maps : Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, 96(6) :2907–2912, 1999.

[TTC01] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9) :5116–5121, April 2001.

[Tve77] A. Tversky. Features of similarity. *Psychological Review*, 84 :327–352, 1977.

[uni] The uniprot database. http ://www.uniprot.org/.

[Ven01] J. Craig Venter. The sequence of the human genome. *Science*, 291(5507) :1304–1351, 2001.

[vR79] C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1979.

[VZVK95] Victor E. Velculescu, Lin Zhang, Bert Vogelstein, and Kenneth W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235) :484–487, 1995.

[WC53] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids ; a structure for deoxyribose nucleic acid. *Nature*, 171(4356) :737–738, April 1953.

[WC05] Wright and Chris C. The kappa statistic in reliability studies : use, interpretation, and sample size requirements. *Physical Therapy*, 85(3) :257–268, 2005.

[WDP+07] James Z. Wang, Zhidian Du, Rapeeporn Payattakool, Philip S. Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10) :1274–1281, 2007.

[Wei01] Alvin M. Weinberg. Messenger rna : origins of a discovery. *Nature*, 414 :485–485, 2001.

[WF99] Ian H. Witten and Eibe Frank. *Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, 1st edition, October 1999.

[WF09] Leland Wilkinson and Michael Friendly. The history of the cluster heat map. *The American Statistician*, 63(2) :179–184, 2009.

[WHJO03] J. Wang, T. Hellem, I. Jonassen, and Others. Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. *BMC Bioinformatics*, 4(60) :1–12, December 2003.

[WHM+09] Nicole L. Washington, Melissa A. Haendel, Christopher J. Mungall, Michael Ashburner, Monte Westerfield, and Suzanna E. Lewis. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol*, 7(11) :e1000247, 11 2009.

[WLB10] Lin Wang, Pinghua Li, and Thomas P. Brutnell. Exploring plant transcriptomes using ultra high-throughput sequencing. *Briefings in Functional Genomics*, 9(2) :118–128, 2010.

[WP94] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

[WSGA03] Chris Wroe, Robert Stevens, Carole A. Goble, and Michael Ashburner. A methodology to migrate the gene ontology to a description logic environment using daml+oil. In *Pacific Symposium on Biocomputing*, pages 624–635, 2003.

[Wu09] Zhijin Wu. A review of statistical methods for preprocessing oligonucleotide microarrays. *Statistical Methods in Medical Research*, 18(6) :533–541, 2009.

[WY05] Kuo-Lung Wu and Miin-Shen Yang. A cluster validity index for fuzzy clustering. *Pattern Recogn. Lett.*, 26(9) :1275–1291, 2005.

[XGZD09] Tao Xu, JianLei Gu, Yan Zhou, and LinFang Du. Improving detection of differentially expressed gene sets by applying cluster enrichment analysis to gene ontology. *BMC Bioinformatics*, 10(1) :240, 2009.

[Yan09] Yunfeng Yang. *Use of Genomic DNA as Reference in DNA Microarrays*, volume 544 of *Methods in Molecular Biology*. June 2009.

[YBDS02] Yee Hwa Yang, Michael J Buckley, Sandrine Dudoit, and Terence P Speed. Comparison of methods for image analysis on cdna microarray data. *Journal of Computational and Graphical Statistics*, 11(1) :108–136, 2002.

[YBS01] Yee Hwa Yang, Michael J. Buckley, and Terence P. Speed. Analysis of cdna microarray images. *Briefings in Bioinformatics*, 2(4) :341–349, 2001.

[YCCK] Ungsik Yu, Yoon Jeong Choi, Jung Kyoon Choi, and Sangsoo Kim. To-go : a java-based gene ontology navigation environment. *Bioinformatics*, 21(17) :3580–3581.

[YG00] Cheng Y and Church GM. Biclustering of gene expression data. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology 2000*, pages 93–103, 2000.

[YHN06] S. Ben Yahia, T. Hamrouni, and E. Mephu Nguifo. Frequent closed itemset based algorithms : a thorough structural and analytical survey. *SIGKDD Explor. Newsl.*, 8 :93–104, June 2006.

[ZBRPB10] Xiangqun Zheng-Bradley, Johan Rung, Helen Parkinson, and Alvis Brazma. Large scale comparison of global gene expression patterns in human and mouse. *Genome Biology*, 11(12) :R124, 2010.

[ZFW+03] Barry R. Zeeberg, Weimin Feng, Geoffrey Wang, May D. Wang, Anthony T. Fojo, Margot Sunshine, Sudarshan Narasimhan, David W. Kane, William C. Reinhold, Samir Lababidi, Kimberly J. Bussey, Joseph Riss, J. Carl Barrett, and John N. Weinstein. Gominer : A resource for biological interpretation of genomic and proteomic data. *Genome Biology*, 4 :28, 2003.

Résumé

L'analyse bioinformatique des données de transcriptomique a pour but d'identifier les gènes qui présentent des variations d'expression entre différentes situations, par exemple entre des échantillons de tissu sain et de tissu malade et de caractériser ces gènes à partir de leurs annotations fonctionnelles. Dans ce travail de thèse, je propose quatre contributions pour la prise en compte des connaissances du domaine dans ces méthodes. Tout d'abord je définis une nouvelle mesure de similarité sémantique et fonctionnelle (IntelliGO) entre les gènes, qui exploite au mieux les annotations fonctionnelles issues de l'ontologie GO ('Gene Ontology'). Je montre ensuite, grâce à une méthodologie d'évaluation rigoureuse, que la mesure IntelliGO est performante pour la classification fonctionnelle des gènes. En troisième contribution je propose une approche différentielle avec affectation floue pour la construction de profils d'expression différentielle (PED). Je définis alors un algorithme d'analyse de recouvrement entre classes fonctionnelles et ensemble des références, ici les PEDs, pour mettre en évidence des gènes ayant à la fois les mêmes variations d'expression et des annotations fonctionnelles similaires. Cette méthode est appliquée à des données expérimentales produites à partir d'échantillons de tissus sains, de tumeur colo-rectale et de lignée cellulaire cancéreuse. Finalement, la mesure de similarité IntelliGO est généralisée à d'autres vocabulaires structurés en graphe acyclique dirigé et enraciné (rDAG) comme l'est l'ontologie GO, avec un exemple d'application concernant la réduction sémantique d'attributs avant la fouille.

Mots-clés: Expression génique, Transcriptome, Exploration de données, Analyse de données symboliques, Syndrome de Lynch, Données transcriptomiques, Cancer colorectal, Ontologie GO, Mesure de similarité sémantique, Graphe acyclique dirigé enraciné, Classification fonctionnelle de gènes, Profils d'expression, Réduction d'attributs.

Abstract

Bioinformatic analyses of transcriptomic data aims to identify genes with variations in their expression level in different tissue samples, for example tissues from healthy versus seek patients, and to characterize these genes on the basis of their functional annotation. In this thesis, I present four contributions for taking into account domain knowledge in these methods. Firstly, I define a new semantic and functional similarity measure which optimally exploits functional annotations from Gene Ontology (GO). Then, I show, thanks to a rigorous evaluation method, that this measure is efficient for the functional classification of genes. In the third contribution, I propose a differential approach with fuzzy assignment for building differential expression profiles (DEPs). I define an algorithm for analyzing overlaps between functional clusters and reference sets such as DEPs here, in order to point out genes that have both similar functional annotation and similar variations in expression. This method is applied to experimental data produced from samples of healthy tissue, colorectal tumor and cancerous cultured cell line. Finally the similarity measure IntelliGO is generalized to another structured vocabulary organized as GO as a rooted directed acyclic graph, with an application concerning the semantic reduction of attributes before mining.

Keywords: Transcriptomic data, Cancer colorectal, Gene ontology, Semantic similarity measure, Rooted directed acyclic graph, Gene functional classification, Expression profile, Dimension reduction.