Le traitement automatique des langues et la fouille des textes en biologie, un nouveau défi pour Gene Ontology

Natalia Grabar, Cédric Bousquet, Marie-Christine Jaulent

Université Paris Descartes, Faculté de Médecine; Inserm, U729; SPIM, 75006 Paris, France
prenom.nom@spim.jussieu.fr

Résumé

Le projet Gene Ontology *(*GO*) a été initié pour proposer un vocabulaire unique pour la description des gènes de différentes espèces. Cette terminologie est utilisée essentiellement pour une annotation manuelle et contrôlée des gènes. Toutefois, face à l'augmentation incessante du nombre de gènes à annoter et du volume de données textuelles à traiter, les applications du traitement automatique des langues et de la fouille de textes sont de plus en plus répandues dans le domaine de la biologie. Mais dans ces applications, l'utilisation des termes de* GO *montre certaines limites, dues entre autres à leur formulation linguistique et à leur structure. Dans cet article, nous nous intéressons à analyser ces limites et à proposer des solutions pour optimiser l'utilisation de* GO *dans les applications automatiques.*

Mots-clés *: Gene Ontology, maintenance et évolution des RTO, synonymie, relations terminologiques, interopérabilité sémantique, biologie, génomique*

Abstract

The Gene Ontology *(*GO*) project has been initiated in order to provide a unique vocabulary common for the description of genes from different species. The terminology is essentially used for the manual and controlled annotation of genes. However, the increase of the number of genes to annotate and the amount of literature to handle, natural language processing applications and text mining are widespread in the biology domain. Nowever, in these applications, the use of* GO *terms shows limits due to their linguistic features and structure. In this paper, we aim at analyzing those limits, and at proposing solutions to optimise the use of* GO *in automatic applications.*

Key-words*: Gene Ontology, maintenance and evolution of terminological and ontological resources, synonymy, terminological relations, semantic interoperability, biology, genomics*

1 INTRODUCTION

La recherche récente en biologie, et en particulier le séquençage des génomes de différentes espèces, a transformé cette science autant dans ses aspects théoriques qu'expérimentaux. De nouvelles connaissances, mises au jour grâce aux travaux de séquençage, ont permis de se rendre compte que les espèces différentes ont des gènes et protéines avec des fonctions similaires et montrent, de ce fait, une conservation fonctionnelle [27]. Par exemple, cette découverte, ou reconnaissance, de la conservation des espèces a permis, ces dernières années, de résoudre quelques cas litigeux de la classification des espèces dans l'arbre phylogénétique. Mais elle conduit surtout vers une unification de la biologie : l'information sur les gènes proches dans divers organismes contribue à une meilleure compréhension de ces organismes. Par exemple, le travail décrit dans [22] a montré qu'environ 12 % des gènes du ver *C. elegans* ont des rôles biologiques qui pourraient être inférés, grâce à la comparaison de séquences de nucléotides avec leurs orthologues pressentis de la levure *S. cerevisiae* (l'orthologie couvre les gènes qui ont des fonctions proches dans des espèces différentes) et à partir de la connaissance déjà accumulée sur les fonctions de ces orthologues. Même si le rôle biologique d'un gène dans un organisme n'est pas identique au rôle du gène similaire dans un autre organisme, cette connaissance permet de faire des inférences sur la vie et le développement de chacun d'entre eux [27]. La découverte de la conservation des espèces a posé la question de la mise à disposition d'un ensemble de termes pour la description des fonctions communes ou proches dans différentes espèces. Ces termes permettraient en effet de mettre en parallèle les fonctions de ces différentes espèces, chose impossible lorsque les biologistes annotent les gènes de différents organismes et créent les vocabulaires nécessaires de manière indépendante. La constatation de cette lacune se trouve à l'origine de *Gene Ontology* (*GO*), conçue comme un outil d'unification de la biologie [27]. Bien qu'ayant rencontré des réticences dans la communauté des ontologies biomédicales, essentiellement à cause de ses principes naïfs et simplistes [45], *GO* a connu et connaît toujours un développement rapide. De par son contenu et son développement, elle s'impose de plus en plus comme un standard dans le domaine de la biologie.

1.1 Annotation fonctionnelle des gènes *vs.* indexation des documents

Le contexte principal d'utilisation de *GO* correspond à l'annotation fonctionnelle des gènes de différentes espèces par leurs fonctions. Cette annotation est le plus souvent faite manuellement à partir de la littérature scientifique du domaine de la biologie. L'annotation fonctionnelle des gènes doit être différenciée de l'annotation des textes, proche de l'indexation.

Dans le cas de l'annotation des textes, un indexeur doit trouver les termes correspondant aux thématiques principales et éventuellement les classer dans

GAIP and RGS4 are GTPase-activating proteins for the Gi subfamily of G protein alpha subunits.

A novel class of regulators of G protein signaling (RGS) proteins has been identified recently. Genetic evidence suggests that RGS proteins inhibit G protein-mediated signaling at the level of the receptor-G protein interaction or the G protein alpha subunit itself. We have found that two RGS family members, GAIP and RGS4, are GTPase-activating proteins (GAPs), accelerating the rate of GTP hydrolysis by Gi alpha 1 at least 40-fold. All Gi subfamily members assayed were substrates for these GAPs ; Gs alpha was not. RGS4 activates the GTPase activity of certain Gi alpha 1 mutants (e.g., R178C), but not others (e.g., Q204L). The GAP activity of RGS proteins is consistent with their proposed role as negative regulators of G protein-mediated signaling.

FIG. 1 – Exemple de résumé.

l'ordre de leur pertinence. Ainsi, dans un article de biologie, les thématiques principales peuvent être traduites par les termes du domaine et/ou les noms de gènes, en fonction des critères de l'indexation et des thésaurus utilisés. Avec l'indexation des textes biologiques, le lien entre les termes décrivant les fonctions (et les gènes) reste implicite. Quant à l'annotation fonctionnelle des gènes, son objectif consiste à détecter dans la littérature les fonctions qui sont expérimentalement reconnues comme étant propres aux gènes étudiés. Ainsi, dans un texte, un annotateur doit reconnaître (1) les noms de gènes, (2) les termes correspondant aux fonctions biologiques et (3) les liens entre les deux. Éventuellement, ces informations peuvent être classées dans l'ordre de leur pertinence. Avec l'annotation fonctionnelle des gènes, le lien entre les gènes et leurs fonctions devient explicite.

La figure 1 présente un exemple de titre et de résumé d'un article de biologie. L'article en question a paru dans la revue *Cell* en 1996. Il est enregistré dans la base bibliographique Medline sous le numéro PMID 8756726. Cet article porte sur les gènes humains *GAIP* et *RGS4*. Les termes MeSH [51], utilisés pour l'indexation des articles de biologie dans cette base, peuvent être des termes chimiques (1) ou non (2). Pour l'indexation de l'article en question, les termes MeSH suivants ont été utilisés (entre parenthèses sont indiqués les qualificateurs des termes) :

1. *DNA, Complementary* ; *G alpha-interacting protein* ; *GTPase-Activating Proteins* ; *Phosphoproteins* ; *Proteins* ; *RGS Proteins* ; *Guanosine Diphosphate* ; *RGS4 protein* ; *Magnesium* ; *Guanosine Triphosphate* ; *GTP Phosphohydrolases* ; *GTP-Binding Proteins*
2. *Base Sequence* ; *DNA, Complementary (analysis)* ; *Escherichia coli* ; *GTP Phosphohydrolases (metabolism)* ; *GTP-Binding Proteins (genetics, metabolism)* ; *GTP-Binding Proteins (genetics - metabolism)* ; *GT-*

Pase-Activating Proteins ; *Guanosine Diphosphate (metabolism)* ; *Guanosine Triphosphate (metabolism)* ; *Humans* ; *Hydrolysis* ; *Magnesium (pharmacology)* ; *Molecular Sequence Data* ; *Mutation* ; *Phosphoproteins (metabolism)* ; *Polymerase Chain Reaction* ; *Proteins (metabolism)* ; *RGS Proteins* ; *Research Support, Non-U.S. Gov't* ; *Research Support, U.S. Gov't, P.H.S.* ; *Temperature*

Certains de ces termes d'indexation apparaissent dans le résumé, d'autres dans l'article complet à partir duquel l'indexation est effectivement réalisée.

Quant à l'annotation fonctionnelle des gènes *GAIP* et *RGS4* à partir du résumé, elle donnerait trois termes *GO* :

- *GTP catabolism* (GO :0006184), également appelé *GTP hydrolysis*
- *GTPase activator activity* (GO :0005096), également appelé *GAP*
- *regulator of G-protein signaling activity* (GO :0016299)

Grâce à ces exemples, nous pouvons voir que les objectifs et les résultats de l'indexation des textes et de l'annotation fonctionnelle de gènes ne sont pas en totale adéquation.

À titre indicatif, notons que dans la version de janvier 2006 de la base SwissProt, le gène *RGS4* est annoté par le terme *signal transducer activity* (GO :0004871), qui est ancêtre du terme *regulator of G-protein signaling activity* (GO :0016299). Quant au gène *GAIP*, il n'est pas annoté dans cette version de Swissprot.

1.2 Dépasser les limitations de l'annotation manuelle des gènes

Dans le domaine de la biologie, ce sont les annotateurs de différentes bases de données, par exemple SwissProt [17], qui effectuent l'annotation fonctionnelle de gènes. Mais l'annotation manuelle ne peut pas suivre l'évolution de la biologie vu le nombre de gènes étudiés et le volume de la littérature scientifique produite. La question se pose autant pour les annotateurs que pour les biologistes « de paillasse » qui effectuent les expériences. Par exemple, suite aux hybridations de l'ADN ou à la comparaison de séquences des gènes, l'interprétation et la validation des résultats met en jeu des centaines voire des milliers de gènes. La consultation des bases de données existantes proposant des informations partielles, les biologistes doivent analyser en plus la littérature scientifique. Dans ce contexte, les méthodes automatiques, comme celles du traitement automatique des langues (TAL) et de la fouille de textes (FT), offrent une perspective intéressante dans l'analyse de gros volumes de données et l'annotation des gènes par leurs fonctions. Ce défi apparaît dans de nombreux projets, par exemple [6, 15, 32], et, plus récemment, lors des compétitions (*TREC Genomics Track*[1] ou *BioCreAtIvE* [5]). Comme il s'agit de proposer des outils d'assistance aux annotateurs et biologistes, il est logique que ces outils soient conçus pour appliquer les

[1] *http ://ir.ohsu.edu/genomics/*

termes *GO*. Il s'agit alors d'un contexte nouveau d'utilisation de *GO*, où ses termes montrent quelques limites vis-à-vis du TAL et de la FT. Pour dépasser à ces limites et rester pertinente dans le nouveau contexte d'exploitation, il est nécessaire que *GO* évolue et soit maintenue à jour.

1.3 Objectifs

L'objectif de notre article consiste à analyser les limites que les termes de *GO* montrent lorsqu'ils sont utilisés dans le nouveau contexte qui fait appel aux méthodes du TAL et de la FT. Nous voulons, en particulier, proposer des solutions pour l'évolution et l'optimisation de cette ressource. Dans la section 2, nous décrivons la nature de *GO* : les objectifs, qui se trouvaient à l'origine de sa conception, et son contenu actuel. Dans la section 3, nous montrons que les limites ressenties lors de l'utilisation des termes *GO* dans les applications automatiques proviennent justement de leur nature et de la manière dont *GO* est construite. Nous présentons et analysons ces limites et proposons quelques évolutions et optimisations effectives et possibles. Nous terminons avec une conclusion (sec. 4).

À travers cet article, nous tâchons de répondre à plusieurs thèmes soulevés par ce numéro spécial de la revue I3 :

- Nous montrons qu'une ressource terminologique est conçue en fonction des besoins premiers auxquels elle doit répondre. Par contre, le changement du paradigme de son utilisation conduit inévitablement à l'émergence de nouveaux besoins qui remettent en question la pertinence de cette ressource. Ainsi, *GO*, conçue pour l'annotation manuelle des gènes par leurs fonctions, doit faire face à une utilisation nouvelle, telle qu'effectuée par des outils automatiques. Pour que la terminologie *GO* reste pertinente et continue d'être utilisée dans le domaine de la biologie moléculaire, elle doit répondre à ces besoins nouveaux et donc à évoluer. Nous développons ce point dans la section 3.
- Plus précisément, nous montrons que les évolutions souhaitables pour une ressource terminologique peuvent être mises au jour grâce à l'étude des textes du domaine. Ces textes, étant écrits par les experts du domaine, fixent le savoir et l'expriment d'une manière assez consensuelle, ce qui assure leur compréhension pas d'autres acteurs du domaine. Ce sont les outils spécifiques proposés par l'ingénierie des connaissances qui permettent d'aborder ces textes et d'en extraire les informations utiles pour l'évolution des ressources terminologiques. Ainsi, dans les sections 3.2 et 3.3, nous montrons l'utilisation effective et possible de tels outils pour le maintien de *GO*.
- Le succès d'utilisation des outils d'ingénierie des connaissances pour la maintenance d'une terminologie dépend bien sûr de leur adéquation aux besoins terminologiques et de leur sensibilité aux connaissances du domaine. Dans la section 3.3, nous tâchons de montrer que les outils de construction des terminologies répondent assez bien aux besoins de la

maintenance des terminologies. En effet, dans les deux situations, les connaissances relevées dans les textes sont modélisées et un compromis doit être trouvé quant à leur représentation. Très souvent, la réussite de cette dernière étape dépend de l'expertise humaine. La connaissance humaine est également nécessaire pour « sensibiliser » les outils d'ingénierie des connaissances aux connaissances spécifiques du domaine d'application. Lors du traitement des textes de biologie moléculaire, ces outils doivent y être adaptés, par exemple, grâce à l'utilisation de ressources lexicales et de corpus spécifiques.

– À travers l'exemple du domaine de biologie moléculaire et de sa terminologie *GO*, nous tâchons de démontrer (dans l'introduction de la section 3) qu'il peut être plus intéressant de maintenir une ressource existante et de l'adapter à un nouveau contexte d'utilisation que de chercher à reconstruire une autre ressource que l'on attend être plus « pertinente ». À notre avis, le fait d'entreprendre une nouvelle initiative de construction de terminologie nécessitera un temps et des efforts certains. Avec le temps, cette terminologie montrera des limites, liées par exemple à sa couverture, à la formulation des libellés de ses termes ou à sa structuration. En effet, ces limites sont « dictées » par la nature du matériel textuel et l'usage de la langue. Avec le changement du contexte d'utilisation, les mêmes limites peuvent apparaître. Nous considérons donc que dans une telle situation, les défis liés à la maintenance d'une ressource terminologique sont très importants et l'emportent sur les défis liés à la construction de nouvelles ressources terminologiques.

2 Le contenu de Gene Ontology

Suite aux récentes évolutions en biologie, le besoin d'un vocabulaire pour la description fonctionnelle des gènes de différentes espèces se faisait ressentir. Le projet Gene Ontology (*GO*) est apparu pour satisfaire ce besoin. Il a commencé en 1998 comme une collaboration entre les bases de données de trois organismes modèles : *FlyBase* (mouches *Drosophila*) [25], *SGD* (levure *Saccharomyces*) [21] et *MGD* (souris *Mouse*) [4]. Depuis, l'utilisation de *GO* s'est répandue à d'autres bases de données (génomes de plantes, d'animaux et de micro-organismes). *GO* poursuit un objectif triple et propose :

1. Un vocabulaire structuré permettant de décrire les fonctions de la biologie moléculaire partagées par des espèces différentes (en novembre 2005, *GO* propose 18 315 termes) ;
2. Des annotations de gènes de ces espèces effectuées au moyen du vocabulaire *GO* par les annotateurs (ou curateurs) des bases de données participantes (le consortium comporte actuellement 14 membres et 4 membres associés) ;
3. Des outils pour le développement et l'utilisation du vocabulaire et des

annotations (de très nombreux outils sont développés autour de *GO*, voir le site *http ://www.geneontology.org/GO.tools.shtml*).

Nous nous intéressons ici à la première partie de l'objectif : constitution d'une terminologie structurée pour la description des fonctions de la biologie moléculaire. Nous analysons la nature de *GO* (sec. 2.1) et de ses termes (sec. 2.2), leurs définitions (sec. 2.3) et structuration (sec. 2.4).

2.1 Une terminologie contrôlée structurée

GO peut être caractérisée comme une terminologie contrôlée structurée. En effet, au stade actuel de son existence, *GO* vise tout d'abord à proposer un vocabulaire nécessaire à l'annotation des gènes et de leurs produits. Ce vocabulaire est structuré par des relations taxinomiques (hiérarchiques et partitives). Pour certains termes, des synonymes sont recensés. Les définitions sont créées pour tous les termes. La construction de *GO* est faite manuellement avec le soucis principal de qualité de cette ressource. Notons néanmoins que de nombreux travaux proposent des évolutions éventuelles à *GO*, par exemple :

- application de la logique de description aux termes de *GO* dans le projet Gene Ontology Next Generation (GONG) [64] ;
- vérification de sa consistance [65] ;
- analyse de ses définitions [43] ;
- réflexion sur les principes de sa structuration [42, 61].

Par ailleurs, *GO* fait partie de l'initiative Open Biomedical Ontologies[2] (OBO), qui supervise plusieurs ressources des domaines biologique et médical. Parmi ces ressources, c'est *GO* qui semble satisfaire au mieux les exigences OBO [43].

2.2 Caractéristique des termes dans Gene Ontology

2.2.1 Création de termes

Les gènes des espèces des bases participantes (*Drosophila* dans *FlyBase*, levure dans *SGD*, etc.) sont annotés avec les termes contrôlés de *GO*. Si, lors de l'annotation de gènes, les curateurs se rendent compte que les notions recherchées n'existent pas, de nouveaux termes peuvent être composés et proposés pour être ajoutés à *GO*. Il existe un certain nombre de conventions quant à leur contenu et forme[3]. L'insertion finale de nouveaux termes doit faire l'objet d'un consensus commun.

[2] *http ://obo.sourceforge.net/*

[3] Voir par exemple la page *www.geneontology.org/GO.usage.shtml*

Contenu des termes

De par leur contenu, les termes doivent correspondre à un des trois arbres hiérarchiques de *GO* : processus biologiques, fonctions moléculaires ou composants cellulaires. Comme le Consortium a pour objectif de décrire des espèces différentes, les termes ne devraient pas être spécifiques à une seule espèce [28]. Mais il est néanmoins possible d'inclure des termes spécifiques. Ceux-ci doivent comporter des précisions sur leur portée, grâce à l'utilisation du modifieur *sensu* (*au sens de*). Les exemples qui suivent, montrent ainsi que le développement embryonnaire est différent selon qu'il s'agit des mammifères ou des insectes :

> *embryonic development (sensu Insecta)* (GO :0001700)
> *embryonic development (sensu Mammalia)* (GO :0001701)

Il existe actuellement plus de 30 modifieurs qui peuvent se positionner à différents niveaux de regroupement des espèces :

> *Insecta*, *Mammalia*, *Eukaryota*, *Vertebrata*, *Drosophila*, ...

où la classe de *Mammalia* appartient à l'embranchement des *Vertebrata*, la famille des *Drosophila* à celui des *Insecta*, et tous les quatre sont des *Eukaryota*. Si un terme comporte la spécification par rapport à une espèce, une famille, etc., ses fils doivent comporter la même spécification.

Forme des termes

Pour la forme des termes, la convention principale porte certainement sur leur nature descriptive, même au risque de redondances lexicales :

> *Aim to be reasonably descriptive, even at the risk of some verbal redundancy. Remember, databases that refer to* GO *terms might list only the finest-level terms associated with a particular gene product. If the parent is* aromatic amino acid family biosynthesis*, then the child should be* aromatic amino acid family biosynthesis, anthranilate pathway*, not just* anthranilate pathway.

Pour la création d'un terme fils, il est conseillé donc de spécifier lexicalement le terme père, c'est-à-dire d'y ajouter les mots qui permettent de le rendre plus précis. Ainsi, dans l'exemple cité, il s'agit des termes :

> *aromatic amino acid family biosynthesis*
> *aromatic amino acid family biosynthesis, anthranilate pathway*

Cette convention de création de termes contribue à la création de termes longs. Quelques chiffres sur leur longueur sont indiqués dans le tableau 1. On voit ainsi que ce sont les termes de 2 à 6 mots qui sont les plus fréquents, avec une nette préférence pour les termes à 3 mots et une longueur moyenne de 3,65 mots par terme. Dans la Snomed [23], une autre terminologie médicale, la longueur moyenne des termes est plus petite : environ 3 mots. Notons que le terme le plus long dans *GO* comporte 29 mots ! Il s'agit d'une fonction moléculaire

oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, NAD or NADH as one donor, and incorporation of two atoms of oxygen into one donor (GO :0016708)

Il existe également des termes de 28, 27, 26, etc. mots. Sur cet exemple nous pouvons voir que les termes *GO*, même étant très longs, sont linguistiquement bien formés.

Nb mots	Nb termes	Nb mots	Nb termes
1	296	5	1 944
2	4 233	6	1 041
3	6 459	7	599
4	3 124	> 7	619

TAB. 1 – Nombre de mots dans les termes de *GO*

D'autres conventions de forme spécifient qu'il est préférable d'utiliser les caractères minuscules, les formes au singulier (à moins que les formes au pluriel soient incontournables), les symboles grecs en toutes lettres latines (*alpha*, *gamma*, etc.), les formes étendues des abréviations. Par ailleurs, il existe un vocabulaire de mots utilisés dans les termes de *GO* (*GODict.DAT*). Il doit être consulté, entre autres, pour l'orthographe. Le vocabulaire *GODict.DAT* contient 10 295 formes. Les termes de *GO* comportant plus de 66 000 formes, chaque forme du vocabulaire est utilisée en moyenne 6,5 fois dans les termes de *GO* (avec la Snomed la moyenne est de 6,1). Cette différence peut être considérée comme significative. Certainement comme une conséquence des instructions données pour la création de nouveaux termes, ces termes présentent une certaine redondance lexicale, parfois vue comme le caractère compositionnel des termes de *GO*.

2.2.2 Compositionnalité des termes

Le principe de compositionnalité spécifie que le sens d'une expression complexe est déterminé par le sens de ses constituants et de règles de leur combinaison [56]. Le sens du terme *inner membrane* (GO :0019866) peut ainsi être dérivé du sens du terme *membrane* (GO :0016020) et du sens de *inner* (*intérieur*). Comme dans cet exemple, ce principe est souvent exprimé au niveau lexical des termes. Et comme, de manière générale, il se vérifie souvent dans les domaines scientifiques et techniques, ce principe est exploité avec succès pour la structuration de termes avec différentes relations : synonymiques [36], hiérarchiques [10, 13, 33] ou transversales [33].

Ce principe se vérifie particulièrement bien dans *GO* et dans la création de ses nouveaux termes, y compris des synonymes, [54, 55]. Dans les exemples suivants, tirés de la hiérarchie des composants cellulaires, l'inclusion lexicale

[41] des termes se présente comme un trait spécifique de leur compositionnalité. Comme nous l'avons noté, la convention de création des termes fils grâce à la spécification des termes père contribue à cette situation :

membrane (GO :0016020)
  inner membrane (GO :0019866)
    mitochondrial inner membrane (GO :0005743)
      mitochondrial inner membrane peptidase complex (GO :0042720)
    plastid inner membrane (GO :0009528)
      chloroplast inner membrane (GO :0009706)
      chromoplast inner membrane (GO :0031899)

L'étude de l'inclusion lexicale des termes *GO* [54] d'une version de 2003 montre que :

– 65,3 % de termes, dont les synonymes, de *GO* contiennent un autre terme de cette même terminologie,
– ces 65,3 % de termes correspondent à 72,2 % de concepts de *GO*.

Ce phénomène touche donc une grande partie des termes *GO*.

Une analyse plus avancée des inclusions montre que la majorité de ces termes se trouvent en relation de subsomption indirecte (plusieurs niveaux hiérarchiques de `is-a`) et la minorité en relation de subsomption directe ou en relation `part-of`[4]. Parmi les compléments de l'inclusion lexicale, par exemple *mitochondrial* dans la paire

mitochondrial inner membrane / *inner membrane*

27,8 % apparaissent dans plus de cinq termes et 54,6 % dans plus de deux. Ces chiffres soulignent le caractère redondant des termes de *GO* et corroborent avec les chiffres sur l'utilisation des mots du vocabulaire *GODict.DAT* présentés auparavant. Les auteurs de cette étude concluent que la formation de nouveaux termes à partir de termes déjà existants est particulièrement présente dans *GO*, que cette formation repose sur un nombre fini de compléments, et que, sans doute, le Consortium encourage la création de nouveaux termes selon ce principe [54].

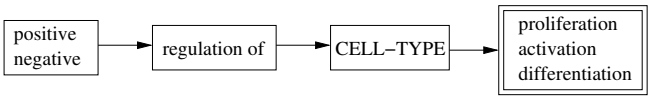

FIG. 2 – Automate pour la description et génération de termes.

Ces redondances de *GO* ne contribuent certainement pas à proposer un ensemble lexicalement représentatif des termes de biologie, ce qui est un point défavorable pour le TAL. Par contre, la compositionnalité peut être utilisée pour la création automatique de nouveaux termes et pour leur validation. Par

[4]La structuration des termes *GO* est présentée plus loin, section 2.4.

exemple, le graphe de la figure 2, proposé dans [55], peut servir pour une génération systématique de très nombreux termes, en sachant que *CELL-TYPE* couvre plusieurs types de cellules (*B-cell*, *lymphocyte*, *epithelial cell*, *urothelial cell*, *fibroblast*, *cell*) :

positive regulation of fibroblast activation
positive regulation of epithelial cell proliferation
negative regulation of epithelial cell proliferation
negative regulation of B-cell differentiation
positive regulation of B-cell activation

Par ailleurs, la fréquence des compléments des termes incluants peut devenir un facteur quant à la validation de nouveaux termes. Ainsi, plus un complément est déjà fréquent dans les termes existants, plus on favorisera la création de nouveaux termes avec ce complément [54].

La compositionnalité peut aussi être utilisée dans la description formelle des termes, comme c'est le cas avec les termes cliniques de la Snomed [62] ou les termes des interventions chirurgicales de la CCAM [57]. Cette perspective demande bien sûr d'avoir déjà formalisé les primitives sémantiques des termes de biologie de même que les règles de leur composition.

2.3 Définitions des termes

En plus d'une liste de termes, *GO* propose également leurs définitions, résultat d'un consensus commun [28]. Pour de nouveaux termes, les annotateurs utilisent, autant que possible, les définitions déja existantes dans *Oxford Dictionary of Molecular Biology*, travaux en biologie moléculaire, base de données sur les protéines SwissProt [2], etc. Si la définition recherchée n'est pas présente dans ces sources, elle est créée par les curateurs de *GO*. Dans tous les cas, l'origine de la définition est enregistrée. Ci-dessous un exemple d'enregistrement pour le terme de processus biologique *positive regulation of fibroblast proliferation* :

id : GO :0048146
name : positive regulation of fibroblast proliferation
namespace : biological process
def : "Any process that activates or increases the rate frequency or extent of multiplication or reproduction of fibroblast cells." (GO :jic)
is-a : GO :0008284 ! positive regulation of cell proliferation
is-a : GO :0048145 ! regulation of fibroblast proliferation

Dans ce format, qui correspond au format *obo* (Open Biomedical Ontologies), partagé par différentes terminologies et ontologies biomédicales, *id* introduit l'identifiant, *name* son libellé, *namespace* la hiérarchie (processus biologique, fonction moléculaire ou composant cellulaire), *def* la définition (dans cet exemple, elle est faite par le curateur *jic* de *GO*), *is-a* la relation hiérarchique (ce terme a deux pères : *positive regulation of cell proliferation* et *regulation of fibroblast proliferation*). Les définitions sont nécessaires à la bonne compréhension des termes et à leur inclusion dans *GO*.

2.4 Structuration des termes

Les termes de *GO* sont structurés en trois arbres (appelés ontologies) : processus biologiques, composants cellulaires et fonctions moléculaires. Ces trois axes de termes ont été choisis parce qu'ils représentent les connaissances nécessaires à la description de la majorité d'organismes [28] et sont donc potentiellement utiles pour l'annotation des gènes de nombreuses espèces.

Les termes sont structurés sous forme d'un graphe dirigé acyclique. Les relations mises en oeuvre sont des relations hiérarchiques ou taxinomiques : relation de subsomption `is-a` et relation partitive `part-of`. La sémantique de ces deux relations s'est précisée au fur et à mesure. Par exemple, la relation `part-of`, qui était vague à l'origine (pouvant signifier *is necessarily part of*, *can be part of* ou *not is always part of*) [61], correspond, dans les versions récentes, à *is necessarily part of*.

Dans la structure *GO*, un terme peut avoir plus d'un père. Par exemple, le terme *aromatic amino acid family biosynthesis* (GO :0009073) a trois pères : *amino acid biosynthesis* (GO :0008652), *aromatic amino acid family metabolism* (GO :0009072) et *aromatic compound biosynthesis* (GO :0019438). Les trois pères sont liés à leur fils *aromatic amino acid family biosynthesis* avec des relations `is-a`. Notons également que les deux relations hiérarchiques `is-a` et `part-of` ne sont pas exclusives. Ainsi, les deux termes père de *mitochondrial inner membrane* (GO :0005743) sont reliés avec lui avec ces deux relations :

`is-a` *inner membrane* (GO :0019866)
`part-of` *mitochondrial membrane* (GO :0005740)

Les termes reçoivent également des synonymes. Plusieurs types de synonymes sont distingués : synonymes exacts, plus spécifiques, plus génériques, viellis et autres. Par exemple, le terme *aromatic amino acid family biosynthesis* (GO :0009073) en a trois :

aromatic amino acid family anabolism
aromatic amino acid family formation
aromatic amino acid family synthesis

Ils sont tous construits sur le même schéma et présentent ici encore une compositionnalité avec la substitution d'un ou plusieurs éléments. Dans cet exemple, il s'agit du paradigme *biosynthesis*, *anabolism*, *formation* et *synthesis*. La compositionnalité des termes se vérifie également lorsqu'ils sont reliés par la relation `is-a` ou `part-of` [54, 55].

Dans la version de novembre 2005, 18 315 termes de *GO* sont reliés avec 24 537 relations `is-a` et 2 726 relations `part-of`. Ils reçoivent 13 850 synonymes.

Si ces trois relations (`is-a`, `part-of` et `synonym`) permettent d'assurer la structuration minimale, et suffisante pour certaines applications, des termes de *GO*, elles sont loin de représenter la variété de relations exis-

tant dans le domaine biomédical. Par exemple, l'UMLS [53] utilise actuellement 54 relations, héritées des terminologies constituantes. En plus des relations hiérarchiques et synonymiques, l'UMLS recense des relations transversales : `located-in`, `transformation-of`, `adjacent-to`, etc, définies pour certaines dans [60]. Si l'effort de conceptualisation des termes de *GO* est fait, certaines de ces relations transversales pourraient servir à relier les primitives sémantiques et à composer les définitions formelles des termes *GO*.

3 UTILISATION DE *GO* DANS DES APPLICATIONS AUTOMATIQUES ET SON ÉVOLUTION FACE À CES NOUVEAUX DÉFIS

Selon les objectifs du Consortium *GO*, cette terminologie a été créée pour l'annotation des gènes et de leurs produits provenant de différentes espèces. L'annotation est effectuée principalement suite à une indexation contrôlée manuelle des gènes et garantit ainsi des données de bonne qualité. Par contre, avec l'augmentation du nombre de gènes à annoter et surtout avec l'augmentation du volume des données scientifiques et expérimentales dans le domaine de la biologie à analyser, les outils d'annotation (semi)automatiques apparaissent comme une alternative intéressante à l'analyse purement manuelle. Mais l'utilisation des termes *GO* dans les tâches d'annotation automatique montre quelques limites qui découlent directement de leurs caractéristiques présentées dans la section précédente :

1. Les libellés des termes *GO* montrent l'utilisation d'un lexique limité, souvent forgé à partir du vocabulaire *GODict.DAT*. La formulation de ces libellés devant être descriptive, même au risque de redondances lexicales, les termes sont, de ce fait, compositionnels et de très nombreux termes fils incluent lexicalement leurs pères. La formulation des termes est lourde et, par conséquent, ils peuvent être difficilement reconnus dans les textes.
2. Même si *GO* évolue en suivant les annotations de différents génomes sa couverture n'est pas exhaustive. D'autres notions et termes sont utilisés par les scientifiques dans les textes mais n'apparaissent pas dans *GO*.
3. La structuration des termes de *GO* est effectuée avec des relations taxinomiques (subsomption `is-a` et partitive `part-of`) et la relation de synonymie. D'autres relations, en particulier les relations transversales entre les termes de différents axes sémantiques, pourraient être utiles. Par exemple, les fonctions moléculaires `sont impliquées` dans les processus biologiques et `ont pour localisation` des composants cellulaires. Le fait de disposer de ces relations peut simplifier

le processus d'annotation, manuel ou automatique, et ouvre des pistes vers une modélisation conceptuelle des termes *GO*.

Bien que *GO* montre des limites, nous pensons qu'il est plus intéressant d'apporter une évolution à ce produit terminologique existant que de développer une nouvelle terminologie de la biologie à partir de documents biomédicaux et/ou des connaissances expertes. Les raisons que nous avançons ici ne sont pas propres au domaine de la biologie mais sont liées à la construction et l'utilisation de tout produit terminologique :

- L'utilisation des termes *GO* permet de suivre l'évolution de cette ressource et d'évoluer avec elle.
- Si *GO* existe et évolue depuis 1998 c'est grâce aux efforts non négligeables de la communauté entière [45] et surtout des annotateurs des bases de données participantes. Le fait d'entreprendre une initiative d'envergure moindre ou similaire à celle de *GO* doit donc prévoir un certain temps nécessaire à la construction d'une nouvelle terminologie, temps où cette terminologie ne sera pas disponible.
- Par ailleurs, la création d'une nouvelle terminologie ne fait que repousser les difficultés vis-à-vis du TAL et de la FT que nous avons mentionnées par rapport à *GO* : couverture, libellés des termes et leur structuration. En effet, les textes de biologie sur lesquels les annotateurs travaillent ne présentent pas des termes identiques.
- En cas d'utilisation d'une terminologie différente de *GO* et lorsqu'une comparaison ou évaluation par rapport aux annotations *GO* enregistrées dans des bases de données est prévue, l'appariement entre les termes *GO* et les termes de cette terminologie devra être effectué.
- Quelle que soit la terminologie utilisée pour l'annotation fonctionnelle des gènes, les termes de cette terminologie devront être appariés avec les expressions utilisées dans les documents traités.

Comme dans d'autres domaines biomédicaux, il s'avère crucial en biologie aussi de développer des ressources et des outils qui permettent d'apparier les termes [11], tout en sachant que dans ce domaine les techniques habituelles ne semblent pas être suffisantes. Cette problématique est d'ailleurs répandue dans la communauté de l'ingénierie des connaissances, comme en témoigne par exemple la campagne OAEI (Ontology Alignment Evaluation Initiative)[5].

Pour une utilisation plus efficace de *GO* dans les applications automatiques du TAL et de la FT, nous voyons plusieurs évolutions interliées :

1. Pour une meilleure reconnaissance des termes *GO* dans les textes, nous avons besoin de détecter leurs synonymes et variantes, accessibles dans des terminologies et lexiques existants ou bien dans les textes (sec. 3.1) ;
2. L'appariement avec d'autres terminologies permet d'améliorer la complétude et surtout la structure des termes *GO* (sec. 3.2) ;

[5] *http ://oaei.ontologymatching.org/2006/*

3. L'enrichissement de la structure des termes se présente comme un objectif en soi (détection de relations transversales entre les termes des trois arbres hiérarchiques), mais aussi comme un moyen pour relier aux termes de *GO* leurs termes synonymes ou bien pour insérer dans les arbres hiérarchiques de nouveaux termes (sec. 3.3).

Ces solutions ressortent pour la plupart des domaines de la terminologie textuelle et de l'ingénierie des connaissances. L'originalité, et surtout la difficulté, viennent du fait de travailler dans le domaine de la biologie et d'avoir à manipuler des notions et des termes spécifiques, dont la maîtrise n'est pas facile à acquérir. Nous sommes intervenus dans le cadre du projet InSerGène et nos contributions sont proposées dans les sections 3.1 et 3.3, au même titre que les contributions d'autres projets.

3.1 Reconnaissance de termes dans les textes

Les termes de *GO* ont une visée descriptive. Leurs libellés doivent être explicites, en comportant l'information générique héritée des pères hiérarchiques et l'information spécifique propre. Ainsi, étant donné le terme père *aromatic amino acid family biosynthesis*, le terme fils doit être *aromatic amino acid family biosynthesis, anthranilate pathway* et pas seulement *anthranilate pathway*. Le fait de forger les termes de cette manière les rend « lourds » du point de vue linguistique. En conséquence, leur projection et reconnaissance dans les textes ne sont pas évidentes. Le Consortium est d'ailleurs conscient de cette situation [18] et se présente aussi comme demandeur de synonymes. Nous illustrons cette situation à travers différentes expériences et en présentons quelques résultats : processus de l'annotation manuelle des fonctions des gènes (sec. 3.1.1), relation entre la longueur des termes et leur reconnaissance (sec. 3.1.2), exploitation de la compositionnalité des termes pour leur reconnaissance (sec. 3.1.3), détection de variantes morphosyntaxiques des termes (sec. 3.1.4). Nous terminons par un bilan (sec. 3.1.5).

3.1.1 Annotation fonctionnelle manuelle des gènes

Dans la pratique, le protocole d'annotation manuelle appliqué par des annotateurs ne permet pas d'effectuer un appariement direct avec les termes *GO* et prévoit une part de leur reconnaissance et sélection, que nous décrivons. Ce protocole est appliqué aux articles complets, sélectionnés à un stade préalable, et imprimés. Les annotateurs, après avoir détecté les parties de ces textes pertinentes pour l'annotation d'un gène, procèdent de la manière suivante [18] :

1. Les annotateurs surlignent les mots et les expressions courtes susceptibles d'être convertis en termes *GO* ;

2. Ces mots et expressions sont utilisés dans le navigateur *GO* DAG-Edit [29] ou l'interface d'interrogation QuickGO [17] pour sélectionner les termes les plus appropriés ;
3. La définition de ces termes est consultée ;
4. Les propositions d'annotations sont faites et envoyées pour accord.

Nous voyons que, déjà au niveau de l'annotation manuelle, il s'agit de trouver les parties éventuellement pertinentes des termes d'annotation des gènes dans les textes et de « composer » le ou les termes nécessaires à l'annotation.

3.1.2 Relation entre la longueur des termes *GO* et leur reconnaissance

Les directives données aux annotateurs des bases de données encouragent la création de termes *GO* longs et lourds. La compétition BioCreAtIvE [5] a montré ainsi qu'il existe une relation entre la longueur de ces termes et leur reconnaissance.

La compétition BioCreAtIvE, dont la dernière a eu lieu en 2005[6], consiste à évaluer les systèmes de fouille de textes et d'extraction d'information à partir des textes de biologie. L'objectif, à terme, est de proposer aux annotateurs des outils qui les aideraient à analyser la littérature et d'en extraire l'information sur les fonctions de gènes, leurs localisations, etc. Les tâches de cette compétition en 2005 sont directement conditionnées par le travail des annotateurs et les données qu'ils ont à manipuler.

Les équipes, qui ont participé à la compétition, pouvaient ainsi réaliser trois tâches liées à l'annotation à partir des articles complets de biologie :

tâche 2.1 : identifier des passages qui contiennent l'information pertinente pour l'annotation des gènes,

tâche 2.2 : extraire les associations entre les gènes et les termes *GO* qui les caractérisent,

tâche 2.3 : parmi tous les articles disponibles, sélectionner ceux qui sont pertinents pour l'annotation de gènes.

À titre d'information, la tâche 1 portait plus spécifiquement sur la reconnaissance des noms de gènes dans les textes. C'est la tâche 2.2 qui est liée le plus directement à l'utilisation des termes *GO* que nous visons : annotation fonctionnelle automatique de gènes. Selon les organisateurs, les résultats d'outils participants sont encourageants mais laissent une grande marge d'évolution [5, 18]. Cette compétition a permis donc d'observer la relation entre la longueur des termes et la précision des annotations de gènes.

D'une manière non surprenante, les termes les plus courts sont les plus faciles à reconnaître. La figure 3 illustre le rapport entre la longueur des termes (*GO term length*) et la précision des annotations extraites (*Percentage TP*) [5]. Ainsi, dans la tâche 2.2, il a été plus facile d'obtenir les annotations correctes des gènes lorsque les termes *GO* sont courts. L'assignation de termes constitués d'un seul mot montre plus de précision que lorsque les termes sont

[6] *http ://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html*

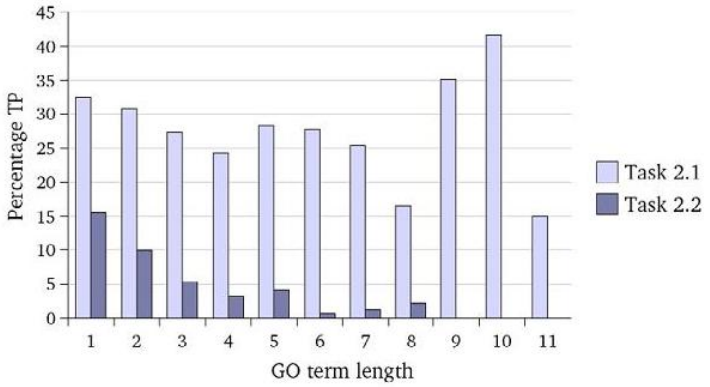

FIG. 3 – Longueur des termes *GO* et la précision des annotations.

composés de 2, 3 ou 4 mots. La précision augmente avec les termes de 5 mots (sans doute l'apport de l'information est plus riche [5]) et diminue avec 6, 7 ou 8 mots. Il n'existe pas d'assignations correctes avec les termes de plus de 8 mots. Il existe donc un rapport inversement proportionnel entre la longueur des termes et la précision des annotations.

Parmi ces résultats, l'assignation des termes de la branche des composants cellulaires montre le plus de précision (34,61 %) et l'assignation des processus biologiques le moins (23,02 %), les fonctions moléculaires étant entre les deux. En effet, les composants cellulaires sont décrits avec les termes les plus courts et les processus avec les plus longs.

Il est donc souhaitable d'utiliser les termes courts dans les applications automatiques ou bien de disposer de règles de leur « recomposition ».

3.1.3 Exploitation de la compositionnalité des termes pour leur reconnaissance

Dans le projet InSerGène d'annotation fonctionnelle des gènes de trois espèces, mouche *D. melanogaster*, ver *C. elegans* et *H. sapiens*, nous tâchons d'exploiter le caractère compositionnel des termes *GO*. Nous travaillons avec les gènes distingués lors des expériences biologiques [44]. Lors de l'annotation de ces gènes, nous appliquons d'abord la mesure de pertinence basée sur le logarithme du facteur de vraisemblance [48] pour calculer les associations entre les gènes et les mots des termes dans les résumés et ensuite nous « recomposons » ces termes [32]. Par exemple, le gène de la mouche *Doa* est associé au terme de processus biologique *sex differentiation* (GO :0007548) parce que les mots de ce terme, *sex* et *differentiation*, sont associés chacun à ce gène. Il en est de même pour le processus biologique *pre-mRNA spli-*

cing (GO :0000398) ou *spindle* (GO :0005819). De la même manière, le gène *18w* est associé au processus *decapentaplegic receptor signaling pathway* (GO :0008101) et le gène *pav* au processus *cytokinesis, contractile ring formation* (GO :0000915).

Lors de l'application de cette méthode, nous n'appliquons pas de règles, logiques ou autres, de recomposition de ces termes. De manière générale, notre approche permet de générer des annotations avec en moyenne 14 % et 15 % de précision et 50 % et 40 % de rappel, lorsque jusqu'à 5 niveaux hiérarchiques de distance sont autorisés par rapport aux annotations de *FlyBase* et *WormBase* respectivement. La comparaison avec les annotations des gènes humains de *SwissProt* n'a pas pu être réalisée à cause de la divergence dans la dénomination des gènes. Bien que les données de travail et la procédure d'évaluation ne soient pas les mêmes, les résultats que nous obtenons sont comparables avec ceux constatés lors de la compétition BioCreAtIvE [5]. Mais notons que nos résultats ne sont pas égaux : nous obtenons entre 0 et 100 % de précision et de rappel selon les gènes. Effectivement, dans certains cas, aucun bon résultat n'est fourni par la méthode, dans d'autres cas, le résultat est proche de la complétude.

Lorsque les annotations ne sont pas validées par les bases de données, cela peut signifier deux choses :

- Les résultats ne sont pas corrects ;
- Les résultats sont nouveaux et non encore enregistrés dans les bases utilisées.

> 1. *pav encodes a kinesin-like protein, PAV-KLP, related to the mammalian MKLP-1. In cellularized embryos, the protein is localized to centrosomes early in mitosis, and to the midbody region of the spindle in late anaphase and telophase.*
> 2. *Doa transcripts and protein are expressed in all cell types both during embryogenesis and in imaginal discs.*

Ainsi, dans l'exemple 1 ci-dessus, le gène *pav* de *D. melanogaster* est associé, parmi autres termes *GO*, aux processus biologiques *anaphase* et *telophase* (GO :0051322 et GO :0051326), mais ces informations sont absentes des fichiers d'annotation de *FlyBase*. Il en est de même pour le gène *Doa* et les processus *imaginal disc development* (GO :0007444) et *embryogenesis* (GO :0009790) de l'exemple 2. Si ces annotations peuvent être difficilement validées par rapport aux annotations des bases de données, une validation manuelle peut distinguer les connaissances fausses de celles qui sont potentiellement intéressantes pour les biologistes. Par contre, il est difficile de mettre en place la validation manuelle. Elle requiert la création d'une interface spécifique et surtout la disponibilité et mobilisation des experts.

L'approche décrite dans cette section permet de contourner la nature linguistique des termes *GO* et recomposer les termes complexes à partir de leurs constituants. Par contre, pour extraire des connaissances plus précises des

textes, il est nécessaire d'améliorer cette méthode et de contraindre sémantiquement la recomposition des termes.

Par ailleurs, des méthodes plus traditionnelles et fondées linguistiquement sont également appliquées pour la reconnaissance des termes *GO*.

3.1.4 Détection de variantes morphosyntaxiques des termes *GO* pour leur reconnaissance

L'objectif du projet CADERIGE [3] ne consiste pas à effectuer l'annotation fonctionnelle des gènes mais à détecter les interactions entre les gènes. Les termes *GO* et ceux provenant d'autres terminologies et lexiques biomédicaux sont utilisés pour améliorer l'analyse syntaxique et les performances du système d'apprentissage. La partie TAL de ce projet donne des comparaisons et indications intéressantes pour l'utilisation des termes *GO*.

Les termes *GO* et ceux provenant d'autres terminologies et lexiques ont été identifiés tels quels et avec la prise en compte de leurs variations morphosyntaxiques traitées par Faster [39]. Cette expérience montre que le repérage de ces variantes permet d'augmenter la reconnaissance des terminologies projetées d'environ 50 % [35]. Elle montre aussi que ce n'est pas *GO* qui permet de couvrir le mieux les corpus constitués à partir des résumés *MedLine*[7] traités : la terminologie MeSH [51] ou même le Glossaire de biochimie et de biologie moléculaire, qui contient moins de 3 000 termes, proposent chacun une meilleure couverture des corpus que *GO*. Cette expérience illustre ainsi qu'il est tout à fait intéressant d'utiliser les variantes morphosyntaxiques des termes *GO*, de même que les termes provenant d'autres terminologies. Dans ce dernier cas, les termes *GO* doivent bien sûr être complétés avec d'autres termes.

Les termes *GO* ont été aussi testés avec les outils TAL d'UMLS [53], entre autres, en vu d'être fusionnés avec d'autres terminologies biomédicales. Dans cette expérience, 79 % des termes ont passé les filtres TAL, ce qui montre que la majorité de ces termes sont bien formés [49]. Mais seulement 35 % parmi eux sont reconnus dans le corpus des résumés *MedLine* et 27 % appariés avec les termes déjà enregistrés dans l'UMLS. Ce travail donne des résultats complémentaires aux précédents quant au taux de reconnaissance des termes *GO* dans les documents *MedLine*. Il montre en particulier que les termes *GO* doivent être complétés avec des termes provenant d'autres terminologies biomédicales.

3.1.5 Bilan

L'ensemble des travaux cités montre que les termes *GO* présentent un potentiel lors de leur utilisation dans des applications du TAL ou de la FT. Par contre, leur utilisation doit être appuyée par des outils spécifiques de détec-

[7] *www.ncbi.nlm.nih.gov/PubMed*

tion de variantes de termes, voire de leur « composition » à partir d'éléments primitifs.

Dans tous les cas, il serait intéressant d'effectuer un recensement systématique des variantes et synonymes des termes *GO*. Si la détection des variantes morphosyntaxiques effectuée avec une méthode, Faster [39], aide à augmenter la reconnaissance des termes *GO* dans un corpus, la détection d'autres types de variantes [34] ou de synonymes [36, 52] permettra d'optimiser encore l'utilisation de *GO*.

Notons que pour l'augmentation de la couverture de *GO* et en particulier pour l'adaptation de son contenu à un corpus, il est possible, en plus du repérage des variantes et synonymes, d'appliquer les méthodes d'acquisition terminologique à partir de corpus [12, 24, 26]. Si cela permet d'enrichir *GO* avec des termes réellement utilisés par les chercheurs, l'effort requis pour leur validation est conséquent. Il n'existe pas vraiment de travaux qui permettraint de quantifier cet effort. [7] considère par exemple qu'avec une interface conviviale il est possible de valider environ 60 termes (du domaine de l'électricité) par heure. Notons que l'expérience montre qu'en général le temps nécessaire à cette tâche est sous-évalué. Par ailleurs, lorsque la validation concerne les relations entre termes, en plus de la validation des termes eux-mêmes, le temps requis sera encore plus important.

Un autre moyen, plus contrôlé, d'enrichissement d'une terminologie, par exemple de *GO*, consiste en sa fusion avec d'autres terminologies. Comme ces autres terminologies sont déjà des ressources contrôlées et modérées, les connaissances qu'elles proposent sont plus sures que celles automatiquement extraites des corpus. Mais il se pose alors à nouveau la question de leur couverture. Dans la section suivante, nous présentons quelques travaux d'appariement de *GO* avec d'autres terminologies.

3.2 Appariement avec d'autres terminologies

Dans cette section, nous mentionnons certains objectifs poursuivis lors de l'appariement des terminologies (sec. 3.2.1) et les approches utilisées pour cet appariement (sec. 3.2.2).

3.2.1 Utilité de l'appariement avec d'autres terminologies

L'appariement entre les termes de différentes terminologies est un champ d'activité productif dans le domaine biomédical, où il existe, en effet, de très nombreuses ressources terminologiques. Par exemple, l'UMLS [53] réunit à ce jour plus de 100 terminologies, classifications et thesaurus. Différents objectifs sont poursuivis lors de l'appariement et fusion des terminologies. Notons par exemple :

– Interopérabilité lexicale et sémantique entre les terminologies. C'est par exemple l'objectif principal d'UMLS [53].

– Enrichissement ou modification de la structuration d'une terminologie grâce à son alignement avec d'autres terminologies, comme par exemple les travaux présentés dans [9, 16].
– Intégration de connaissances nouvelles. Par exemple, l'intégration d'informations absentes dans *GO* : description anatomique spécifique à une espèce [38].
– Détection des synonymes et variantes des termes.

Le dernier objectif apparaît également comme un moyen de réalisation des trois premiers.

3.2.2 Approches pour l'appariement avec d'autres terminologies

Parmi les approches utilisées pour l'appariement de termes, les approches lexicales d'UMLS ont permis de relier 27 % de termes *GO* aux concepts existants dans l'UMLS [49]. Aujourd'hui, cela couvre environ 18 % des termes *GO*. L'approche utilisée dans l'UMLS fait appel aux méthodes qui neutralisent la variation des termes au niveau de la casse, de la ponctuation, de l'ordre de mots et de variations morphologiques en fonction des dictionnaires encodés dans le Specialist Lexicon d'UMLS. Notons que ces méthodes sont proches des travaux décrits dans la section 3.1.4.

Les méthodes dites « hybrides » [19], combinent les approches lexicales d'UMLS [49], la synonymie et les méthodes de recherche d'information. Elles permettent d'augmenter le nombre des appariements avec les termes *GO*. Enfin, lorsque les termes *GO* doivent être dépassés, l'appariement peut aussi être effectué à travers leurs définitions [40], ce qui améliore également les résultats d'appariement avec les termes *GO*. Ce dernier travail a ainsi permis de générer plus de 90 000 nouvelles relations. L'évaluation manuelle a montré que 1 926 relations sur 2 389 analysées sont correctes.

3.3 Enrichissement de la structuration de *GO*

Cette section est consacrée spécifiquement à l'évolution de la structuration des termes *GO*. Nous rappelons d'abord les principes de la structuration actuelle des termes *GO* (sec. 3.3.1), nous présentons ensuite l'utilité de son évolution et, en particulier, de la détection de relations transversales (sec. 3.3.2), et les approches qui permettent de faire évoluer la structure des termes *GO* (sec. 3.3.3).

3.3.1 Structuration actuelle des termes *GO*

Les termes *GO* sont structurés en trois hiérarchies séparées : fonctions moléculaires, processus biologiques et localisations cellulaires. La structuration au sein des hiérarchies est effectuée avec deux types de relations : `is-a` et `part-of`. Certains des termes reçoivent des synonymes. Mais, tandis que les termes de ces trois hiérarchies sont étroitement liés entre eux du point

de vue biologique, ce fait est actuellement absent de *GO*. Le lien entre les termes de ces différentes hiérarchies peut être réalisé avec des relations transversales.

3.3.2 Utilité des relations transversales

Les relations transversales peuvent être utilisées de différentes manières. Principalement, elles permettent de décrire les connaissances d'un domaine plus finement.

Ainsi, dans la base de données *FlyBase*, le fichier *FBgn.summary.acode*, qui contient les annotations de gènes, applique le schéma suivant de description des gènes :

> *A gene encodes a product with the molecular function involved in the biological process and localized to the cellular component.*
>
> *(Un gène encode un produit qui a la fonction moléculaire, impliqué dans le processus biologique et localisé dans un composant cellulaire.)*

Comme exemple de l'application de ce schéma :

> *D. melanogaster gene* strawberry notch *encodes a product with putative ATP binding involved in imaginal disc morphogenesis which is localized to the nucleus ...*

En effet, les annotations faites d'après ce schéma offrent des informations complètes sur un gène. Elles utilisent pleinement le potentiel actuel de *GO*.

Les relations transversales peuvent aussi accomplir un rôle prédictif. Il apparaît par exemple, qu'il est possible de prédire les localisations sous-cellulaires des gènes lorsque leurs fonctions moléculaires sont connues [47]. De la même manière, les annotateurs de *GO* utilisent souvent les associations les plus fréquentes entre les termes des gènes déjà annotés pour proposer des annotations plus complètes pour un nouveau gène [18].

Une structuration plus complète des termes permet en outre d'effectuer des calculs de proximité sémantique entre ces termes [46, 14] et de faire évoluer son modèle conceptuel.

Pour faciliter la description de gènes d'après le schéma mentionné plus haut, ou pour utiliser les relations transversales autrement, il serait intéressant de détecter systématiquement ces relations entre les termes des trois hiérarchies de *GO*.

3.3.3 Approches pour la détection de relations et l'évolution de la structuration de *GO*

Pour la détection de relations transversales entre les termes *GO* plusieurs approches sont possibles. Certaines d'entre elles profitent des données déjà existantes et validées, d'autres chercheront à les acquérir à partir de textes.

L'ensemble de ces approches n'est pas spécifique à la biologie, mais les données à partir desquelles l'acquisition est effectuée le sont.

Acquisition à partir des données contrôlées et validées

Le résultat de l'indexation des articles de la base bibliographique *MedLine* avec les mots clés MeSH [51] peut ainsi être utilisé afin de calculer les associations entre ces différents termes de l'indexation [8, 1]. À leur tour, les associations peuvent conduire vers des relations transversales. Comme l'indexation dans *MedLine* est effectuée avec les termes MeSH, il faut les apparier avec les termes *GO* pour induire des relations entre ces derniers. Cet appariement est possible grâce à l'UMLS où ces deux terminologies, parmi d'autres, sont fusionnées. Mais, comme nous l'avons noté dans la section 3.2.2, environ 18 % des termes *GO* sont appariés. Les travaux qui exploitent l'indexation *MedLine* peuvent donc couvrir au mieux cet ensemble de termes *GO*.

L'annotation fonctionnelle des gènes dans les bases de données peut également être exploitée. Les associations entre termes sont par exemple utilisées pour aider à l'annotation manuelle des fonctions de gènes [18]. Ces associations pourraient de plus être entérinées dans la terminologie [1].

Acquisition à partir des données textuelles

En cas d'acquisition de relations à partir de corpus textuels des méthodes en contexte et hors contexte sont possibles [30].

Détection de relations en contexte. Parmi les méthodes d'acquisition de relation en contexte, [30] distinguent des méthodes de classification de termes et des méthodes à base de marqueurs et patrons lexico-syntaxiques.

Parmi les *méthodes de classification*, mentionnons les règles d'association [63], la mesure de pertinence basée sur le logarithme du facteur de vraisemblance [32] et l'approche distributionnelle [37]. Les deux premières sont appliquées dans le projet InSerGène pour calculer les associations entre les gènes et les termes *GO*. Mais elles peuvent renseigner en même temps sur les relations entre les termes *GO*. Voyons par exemple ces règles d'association générées par [63] :

$foxo_g$ => *response to oxidative* $stress_{pb}$, $transcription_{pb}$, $cell_{cc}$
fas_g => $apoptosis_{pb}$, $death_{pb}$, $signaling_{pb}$, $ligand_{fm}$, *tumor necrosis factor* $_{fm}$
atpase activity $_{fm}$ => $phosphorylation_{pb}$, $microfilaments_{cc}$

Dans cet exemple, *pb* étiquette les processus biologiques, *fm* les fonctions moléculaires, *cc* les composants cellulaires, *g* les gènes et protéines. À travers ces associations, le gène *foxo* est lié avec les termes *GO response to*

We conclude that CDC27 and CDC16 are evolutionarily conserved components of the centrosome and mitotic spindle that control the onset of postmetaphase events `during` *mitosis.*

Eph receptor tyrosine kinases (RTK) and their ephrin ligands are involved in the transmission of signals which regulate cytoskeletal organisation and cell migration, and are expressed in spatially restricted patterns at discrete phases `during` *embryogenesis.*

In vivo transcription factor recruitment `during` *thyroid hormone receptor-mediated activation.*

FIG. 4 – Exemples de phrases avec la marqueur `during`.

oxidative stress, *transcription* et *cell*. Ces mêmes règles impliquent également qu'il existe des relations transversales entre les processus biologiques, les composants cellulaires et les fonctions moléculaires :

response to oxidative stress$_{pb}$ / *cell*$_{cc}$
transcription$_{pb}$ / *cell*$_{cc}$
apoptosis$_{pb}$ / *tumor necrosis factor*$_{fm}$
death$_{pb}$ / *tumor necrosis factor*$_{fm}$

Contrairement aux travaux qui utilisent les termes de l'indexation MeSH de la base *MedLine* [8, 1], les relations transversales présentées ici sont acquises directement des textes. Elles présentent des données intéressantes pour une évolution de la structure des termes *GO*.

Quant aux *méthodes à bases de marqueurs et patrons*, elles se repèrent par rapport aux ancres lexicales et syntaxiques. La figure 4 propose ainsi quelques exemples d'utilisation du marqueur `during`. Ce marqueur encode le plus souvent des relations temporelles. La reconnaissance de ce marqueur et des termes *GO* montre qu'il existe des relations de temporalité entre les termes de différents axes de *GO* ou bien entre les termes d'un même axe :

1. *centrosome*$_{cc}$ / *mitosis*$_{bp}$
2. *spindle*$_{cc}$ / *mitosis*$_{bp}$
3. *transcription factor*$_{bp}$ / *thyroid hormone receptor-mediated activation*$_{mf}$
4. *cytoskeleton organisation*$_{bp}$ / *embryogenesis*$_{bp}$

Dans les deux premiers couples de termes, qui relient les composants cellulaires aux processus biologiques, il s'agit plutôt de relations de localisation. En effet, le processus *mitosis* est assuré par les gènes localisés dans les composants cellulaires indiqués. Quant au couple (3), il montre qu'un processus biologique peut être composé d'une ou de plusieurs fonctions moléculaires. De manière similaire, selon l'exemple (4), un processus biologique complexe peut être constitué de plusieurs processus plus simples. Une telle relation entre la temporalité et la relation `part-of` a déjà pu être remarquée

[58]. L'ensemble des relations pouvant être détectées avec les marqueurs et patrons sont intéressantes pour l'évolution de la structuration de *GO*. Pour certaines relations (hiérarchiques, partitives, synonymes) les patrons et marqueurs sont déjà connus et ont été exploités dans d'autres domaines [59, 50]. Mais ils doivent être adaptés au domaine de la biologie. Par ailleurs, des patrons et marqueurs spécifiques à ce domaine doivent être relevés et interprétés.

Détection de relations hors contexte. Les méthodes appliquables hors contexte s'appuient sur l'étude de la structure des termes. On distingue alors l'application des règles de transformation [39, 36] et la détection des inclusions lexicales [10, 13, 33]. Une des méthodes à base de règles de transformation [39] est appliquée aux termes *GO* pour la détection de leurs variantes dans les documents [35]. Quant à d'autres méthodes, elles doivent encore être adaptées à ce domaine.

Dans la section 3.1.3, nous avons par ailleurs indiqué que les termes *GO* ont une nature compositionnelle. Cette constatation peut également être exploitée pour leur structuration, par exemple à travers l'utilisation de certains marqueurs lexicaux. Nous remarquons ainsi la spécificité de marqueurs selon les arbres hiérarchiques des termes *GO* :

- processus biologiques : *regulation*, *metabolism*, *biosynthesis*, *catabolism* ;
- fonctions moléculaires : *activity*, *binding*, *acting*, *receptor* ;
- composants cellulaires : *complex*, *protein*, *membrane*, *vesicle*, *organelle*.

Cette constatation étant faite, ces mêmes marqueurs peuvent indiquer à laquelle des hiérarchies appartient un terme. Par exemple, si un terme nouveau contient le marqueur *activity*, il pourra être assigné à la hiérarchie des fonctions moléculaires. Dans ce cas, le terme ne sera peut-être pas toujours positionné dans un endroit précis de la hiérarchie des termes. Mais il pourra au moins être typé sémantiquement, ce qui facilitera son incorporation dans la terminologie. [54] remarque par ailleurs que certains de ces marqueurs pourraient même indiquer la nature, `is-a` ou `part-of`, de la relation entre les termes.

La majorité des approches pour la détection de relations hors contexte entre les termes *GO* doit être adaptée et appliquée au domaine de la biologie. Comme d'autres approches citées, elles présentent un potentiel intéressant pour l'évolution de la structuration des termes *GO*. Notons aussi que la combinaison de ces différentes approches, en et hors contexte, donnerait des résultats plus fiables et complets [31].

4 CONCLUSION

Dans cet article, nous avons présenté *Gene Ontology* (*GO*), une terminologie de biologie moléculaire, qui a pour objectif de proposer un vocabulaire

nécessaire à l'annotation fonctionnelle de gènes provenant de différentes espèces. Actuellement, de par son contenu et son développement, elle s'impose comme un standard dans le domaine de biologie moléculaire. Le contexte principal d'utilisation de cette terminologie correspond à l'annotation des gènes par leurs fonctions, telle qu'effectuée par les annotateurs des bases de données biologiques et les biologistes « de paillasse ». Elle est utilisée essentiellement lors du processus de l'annotation manuelle. Avec l'augmentation du nombre de gènes à annoter et du volume des données biologiques à analyser, les outils d'annotation automatique apparaissent comme une alternative intéressante. Ayant pour objectif de proposer de tels outils, nous avons cherché à utiliser les termes *GO* à travers les outils d'annotation automatique de gènes, et nous nous sommes très vite rendus compte que ces termes montrent alors quelques limites. Citons en particulier la formulation lourde des termes, leur caractère compositionnel, leur longueur et leur structuration. Nous avons donc analysé plusieurs de ces limites et proposé des méthodes pour l'évolution et l'optimisation de l'usage de *GO*.

Nous mentionnons en particulier des méthodes désignées pour la reconnaissance des variantes de termes *GO* dans les articles scientifiques et enrichissement de leur structuration. De manière générale, l'utilisation de plusieurs de ces méthodes dans le domaine de biologie moléculaire a immérgé de nos expériences menées dans le cadre du projet InSerGène, où nous intervenons afin de proposer des outils pour l'annotation automatique des gènes par leurs fonctions. Nous résumons ici l'utilisation de quelques unes de ces méthodes.

Parmi les méthodes de reconnaissance des termes *GO*, une perspective intéressante semble correspondre à l'exploitation de leur nature compositionnelle. Ainsi, si les parties constituantes d'un terme sont associées avec un gène donné, il est possible d'induire la relation d'association entre ce gène et le terme entier. En d'autres mots, il est alors possible de proposer l'annotation du gène par la fonction correspondant à ce terme « recomposé » [32]. Notons aussi que cette démarche rappelle le protocole utilisé par les annotateurs dans le processus de l'annotation manuelle. Par ailleurs, la détection de synonymes des termes *GO* à travers une approche proposée par [36], qui exploite également la compositionnalité des termes, est intéressante et prometteuse. Cette approche est actuellement en cours d'adaptation à la langue anglaise et au domaine de biologie. Elle pourra bientôt être testée avec les termes de *GO* sur un corpus d'articles scientifiques.

D'autres approches, actuellement utilisées pour l'annotation fonctionnelle des gènes dans le cadre d'InSerGène, qui utilisent les algorithmes de classification [63, 20], offrent des associations qui conduisent, d'une part, vers l'annotation des gènes et, d'autre part, vers des propositions d'une structuration plus fine des termes *GO*. Les deux aspects sont actuellement étudiés. Notons aussi, que la mise au jour de patrons lexico-syntaxiques, qui permetteraient d'étudier les relations temporelles entre les processus ou bien de détecter le lien entre les processus biologiques complexes et des fonctions

moléculaires, est également en cours. Cette connaissance contribuera à une structuration plus fine de *GO*.

L'ensemble des méthodes appliquées aux termes *GO*, qu'il s'agisse de la détection de ces termes dans les articles scientifiques ou de l'enrichissement de leur structuration, constribue à la maintenance et à l'évolution de cette ressource terminologique dans le temps et dans le contexte de son utilisation dans un nouveau paradigme. En effet, les outils automatiques demandent à disposer de ressources plus exhaustives et bien construites pour que leur utilisation soit optimale. À travers les quelques exemples cités dans cet article, il apparaît par ailleurs que les termes de biologie présentent un matériel spécifique et que la validation des données générées par les différents travaux présentés nécessite une intervention des experts en biologie de plusieurs espèces, ce qui correspond à un travail loin d'être négligeable.

Le travail présenté dans cet article tire un grand profit de l'expérience accumulée au sein de la discipline de l'ingénierie des connaissances. En effet, la majorité des méthodes destinées à faire évoluer le contenu et la structure de *GO*, que nous mentionnons, proviennent de cette discipline. En même temps, l'expérience de maintenance de *GO* n'est pas spécifique à cette terminologie ni au domaine de biologie moléculaire, mais contribue à consolider l'état de l'art de l'ingénierie des connaissances en général.

D'une part, nous avons montré que les méthodes déjà appliquées à d'autres domaines scientifiques et techniques peuvent également être utilisées dans le domaine de la biologie moléculaire. Cette utilisation requiert toutefois leur adaptation au matériel spécifique de ce domaine, par exemple grâce aux ressources lexicales ou corpus spécifiques, et à l'implication conséquente des experts.

D'autre part, indépendamment des domaines d'application, le cycle de vie des produits terminologiques est similaire, surtout en ce qui concerne le besoin de leur maintenance vis-à-vis de l'évolution dans le temps et de leur utilisation dans des contextes nouveaux ou différents. Il s'agit entre autre, de pouvoir faire face à l'évolution de la langue et de la connaissance dans un domaine de spécialité. Nous avons alors tâché de montrer que les outils de construction des ressources terminologiques peuvent être utilisés pour la maintenance de ressources déjà existantes. À notre avis, qu'il s'agisse de la construction ou de la maintenance des terminologies, les connaissances relevées dans les textes sont modélisées en fonction du compromis lié à leur représentation. Très souvent, le caractère consensuel et la réussite de ce compromis nécessitent une implication de l'expertise humaine et en dépendent. Finalement, nous considérons que dans l'état actuel de l'art en ingénierie des connaissances, une réflexion sur la maintenance et évolution des ressources existantes s'avère être plus pertinente que la satisfaction des défis liés à la construction de nouvelles ressources. Ceci surtout si le domaine de spécialité concerné dispose déjà de ressources terminologiques et si les contextes nouveaux d'utilisation restent proches des contextes déjà explorés. Par ailleurs, face aux progrès et à l'évolution des connaissances dans plusieurs domaines

de spécialité, il semble que le besoin de mise à jour, d'enrichissement, de maintenance ou d'evolution des ressources terminologiques, sera de plus en plus présent.

Remerciements

Ce travail a été financé par le réseau d'excellence *Semantic Mining* de la Communauté Européenne. Les expériences décrites s'inscrivent dans le WP24 *Text mining* et dans la plateforme *Biologie des systèmes* de Paris 5. Les auteurs remercient Thierry Hamon et le comité scientifique de ce numéro pour les relectures pertinentes et attentives.

RÉFÉRENCES

[1] Marc Aubry, Annabelle Monnier, Céline Chicault, Marie de Tayrac, Marie-Dominique Galibert, Anita Burgun et Jean Mosser. Combining evidence, biomedical literature and statistical dependence : new insights for functional annotation of gene sets. *BMC Bioinformatics*, 7(241), 2006.

[2] A Bairoch et R Apweiler. The Swiss-Prot protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Research*, 24 :21–25, 1996.

[3] Philippe Bessières, Adeline Nazarenko et Claire Nédellec. Apport de l'apprentissage à l'extraction d'information : le problème de l'identification d'interactions géniques. In *CIDE 2001, 4e Colloque International sur le Document Electronique*, Toulouse, 2001.

[4] JA Blake, JT Eppig, JE Richardson et MT Davisson. The mouse genome database (MGD) : a community resource. status and enhancements. the mouse genome informatics group. *Nucl. Acids Res*, 26 :130–137, 1998.

[5] Christian Blaschke, Eduardo Andres Leon, Martin Krallinger et Alfonso Valencia. Evaluation of biocreative assessment of task 2. *BMC Bioinformatics*, 6(Suppl 1) :S16, 2005.

[6] Christian Blaschke, Juan C. Oliveros et Alfonso Valencia. Mining functional information associated with expression arrays. *Functional & Integrative Genomics*, 1(4) :256–268, 2001.

[7] Henri Boccon-Gibod. Terminology and industrial applications. In URI INIST CNRS, éditeur, *Terminologie et Intelligence artificielle (TIA)*, Nancy, 2001. Invited lectures.

[8] Olivier Bodenreider, Marc Aubry et Anita Burgun. Non-lexical approaches to identifying associative relations in the gene ontology. In *Pacific Symposium of Biocomputing*, pages 91–102, 2005.

[9] Olivier Bodenreider et Anita Burgun. Linking the gene ontology to other biological ontologies. In *ISBM Bio-ontologies SIG meeting*, 2005.

[10] Olivier Bodenreider, Anita Burgun et Thomas C. Rindflesch. Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. In URI INIST CNRS, éditeur, *Terminologie et Intelligence artificielle (TIA)*, pages 11–21, Nancy, 2001.

[11] Olivier Bodenreider, Joyce A. Mitchell et Aleca T. McCray. Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics. In *Annual Symposium of the American Medical Informatics Association (AMIA)*, pages 61–65, 2002.

[12] Didier Bourigault. Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *Traitement Automatique des Langues (TAL)*, pages 105–117, 1993.

[13] Didier Bourigault. Extraction et structuration automatiques de terminologie pour l'aide à l'acquisition de connaissances à partir de textes. In *Reconnaissance de Formes et Intelligence Artificielle (RFIA)*, pages 1123–1132, 1994.

[14] Cédric Bousquet. Mesures de distances dans gene ontology. un projet de comparaison avec blast. Technical report, Université de Pharmacie Paris V, 2004.

[15] EI Boyle, S Weng, J Gollub, H Jin, D Botstein, JM Cherry et G 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.

[16] Anita Burgun et Olivier Bodenreider. An ontology of chemical entities helps identify dependence relations among gene ontology terms. In *Semantic mining in biomedicine*, 2005.

[17] Evelyn Camon, Michele Magrane, Daniel Barrell, Vivial Lee, Emily Dimmer, John Maslen, David Binns, Nicola Harte, Rodrigo Lopez et Rolf Apweiler. The Gene Ontology Annotation (GOA) database : sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, 32(Database issue) :D262–6, 2004.

[18] Evelyn B Camon, Daniel G Barrell, Emily C Dimmer, Vivian Lee, Michele Magrane, John Maslen, David Binns et Rolf Apweiler. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics*, 6(Suppl) :S17, 2005.

[19] M. N. Cantor, I. N. Sarkat, R. Gelman, F. Hartel, O. Bodenreider et Y. A. Lussier. An evaluation of hybrid methods for matching biomedical terminologies : mapping the Gene Ontology to the UMLS. In *Medical Informatics in Europe (MIE)*, pages 62–7, Saint-Malo, France, 2003.

[20] Valentina Ceausu et Sylvie Després. Text mining supported terminology construction. In *5th International Conference on Knowledge Management I-KNOW 2005*, Graz, Austria, 2005.

[21] JM Cherry, C Adler, C Ball, SA Chervitz, SS Dwight, ET Hester, Y Jia, G Juvik, T Roe, M Schroeder, S Weng et D Botstein. SGD : Saccharomyces genome database. *Nucl. Acids Res*, 26 :73–79, 1998.

[22] SA Chervitz, ET Hester, CA Ball, K Dolinski, SS Dwight, MA Harris, G Juvik, A Malekian, S Roberts, T Roe, C Scafe, M Schroeder, G Sherlock, S Weng, Y Zhu, JM Cherry et D Botstein. Using the saccharomyces genome database (SGD) for analysis of protein similarities and structure. *Nucleic Acids Res*, 27(1) :74–78, 1999.

[23] Roger A. Côté, Louise Brochu et Lyne Cabana. *SNOMED Internationale – Répertoire d'anatomie pathologique*. Secrétariat francophone international de nomenclature médicale, Sherbrooke, Québec, 1997.

[24] Béatrice Daille. Repérage et extraction de terminologie par une approche mixte statistique et linguistique. *Traitement Automatique des Langues (T.A.L.)*, 36(1-2) :101–118, 1995.

[25] FlyBase Consortium. The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Research*, 22(17) :3456–3458, 1994.

[26] Yueyu Fu, Travis Bauer, Javed Mostafa, Mathew Palakal et Snehasis Mukhopadhyay. Concept extraction and association from cancer literature. In *WIDM '02 : Proceedings of the 4th international workshop on Web information and data management*, pages 100–103, New York, NY, USA, 2002. ACM Press.

[27] Gene Ontology Consortium. Gene Ontology : tool for the unification of biology. *Nature genetics*, 25 :25–29, 2000.

[28] Gene Ontology Consortium. Creating the Gene Ontology resource : design and implementation. *Genome Research*, 11 :1425–1433, 2001.

[29] Gene Ontology Consortium. The gene ontology (go) database and informatics resources. *Nucleic Acids Res*, 32(Database) :D262–266, 2004.

[30] Natalia Grabar et Thierry Hamon. Les relations dans les terminologies structurées : de la théorie à la pratique. *Revue d'Intelligence Artificielle (RIA)*, 18(1), 2004.

[31] Natalia Grabar et Blandine Jeannin. Contribution de différents outils à la construction d'une terminologie pour la recherche d'information. In Catherine Gréboval, éditeur, *Ingénierie des connaissances (IC)*, Rouen, 2002. Poster.

[32] Natalia Grabar, Magali Sillam, Marie-Christine Jaulent, Céline Lefebvre, Édouard Henrion et Christian Néri. From likelihoodness between words to the finding of functional profile for ortholog genes. In

RANLP 2005 WS on Text Mining Research, Practice and Opportunities, pages 49–55, Borovets, Bulgaria, 2005.

[33] Natalia Grabar et Pierre Zweigenbaum. Lexically-based terminology structuring. In *Terminology*, volume 10, pages 23–54, 2004.

[34] Natalia Grabar, Pierre Zweigenbaum, Lina Soualmia et Stéfan J. Darmoni. Matching controlled vocabulary words. *Studies in Health Technology and Informatics*, 95 :445–450, 2003.

[35] Thierry Hamon. Indexer les documents spécialisés : les ressources terminologiques contrôlées sont-elles suffisantes ? In *Terminologie et Intelligence artificielle (TIA)*, Rouen, 2005.

[36] Thierry Hamon, Adeline Nazarenko et Cécile Gros. A step towards the detection of semantic variants of terms in technical documents. In *International Conference on Computational Linguistics (COLING-ACL'98)*, pages 498–504, Université de Montréal, Montréal, Quebec, Canada, 1998.

[37] Z. S. Harris. *Structures mathématiques du langage*. Monographies de linguistique mathématique. Dunod, Paris, 1971. Traduit par C. Fuchs.

[38] David P Hill, Judith A Blake, Joel E Richardson et Martin Ringwald. Extension and integration of the Gene Ontology (GO) : combining GO vocabularies with external vocabularies. *Genome research*, 12 :1982–1991, 2002.

[39] Christian Jacquemin. A symbolic and surgical acquisition of terms through variation. In S. Wermter, E. Riloff et G. Scheler, éditeurs, *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438, Springer, 1996.

[40] Helen L. Johnson, K Bretonnel Cohen, William A Baumgartner, Zhiyong Lu, Michael Bada, Todd Kester, Hyunmin Kim et Lawrence Hunter. Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. In *Pacific Symposium of Biocomputing*, pages 28–39, 2006.

[41] Georges Kleiber et Irène Tamba. L'hyperonymie revisitée : inclusion et hiérarchie. *Langages*, 98 :7–32, juin 1990. L'hyponymie et l'hyperonymie (dir. Marie-Françoise Mortureux).

[42] Anand Kumar et Barry Smith. *The Unified Medical Language System and the Gene Ontology : Some critical reflections*. Springer, 2003.

[43] Jacob Köhler, Katherine Munn, Alexander Ruegg, Andre Skusa et Barry Smith. Quality control for terms and definitions in ontologies and taxonomies. *BMC Bioinformatics*, 7(212), 2006.

[44] Céline Lefebvre, Jean-Christophe Aude, Éric Clément et Christian Néri. Balancing protein similarity and gene co-expression reveals new links between genetic conservation and developmental diversity in invertebrates. *Bioinformatics*, 21(8) :1550–1558, 2005.

[45] Suzanna E Lewis. Gene Ontology : looking backwards and forwards. *Genome Biology*, 6(1) :103, 2005.

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

[47] Z. Lu et L. Hunter. GO molecular function terms are predictive of subcellular localization. In *Pacific Symposium of Biocomputing*, pages 151–161, 2005.

[48] C. D. Manning et H. Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, 1999.

[49] Alexa T McCray, Allen C Browne et Olivier Bodenreider. The lexical properties of the Gene Ontology (GO). In *Annual Symposium of the American Medical Informatics Association (AMIA)*, pages 504–508, 2002.

[50] Emmanuel Morin. Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique. *Traitement Automatique des Langues (TAL)*, 40(1) :143–166, 1999.

[51] National Library of Medicine, Bethesda, Maryland. *Medical Subject Headings*, 2001. `http://www.nlm.nih.gov/mesh/meshhome.html`.

[52] Adeline Nazarenko, Pierre Zweigenbaum, Benoît Habert et Jacques Bouaud. Corpus-based extension of a terminological semantic lexicon. In *Recent Advances in Computational Terminology*, pages 327–351. John Benjamins, 2001.

[53] NLM. *UMLS Knowledge Sources Manual*. National Library of Medicine, Bethesda, Maryland, 2005. www.nlm.nih.gov/research/umls/.

[54] PV Ogren, KB Cohen, GK Acquaah-Mensah, J Eberlein et L Hunter. The compositional structure of Gene Ontology terms. In *Pacific Symposium of Biocomputing*, pages 214–225, 2004.

[55] PV Ogren, KB Cohen et L Hunter. Implications of compositionality in the Gene Ontology for its curation and usage. In *Pacific Symposium of Biocomputing*, pages 174–185, 2005.

[56] Barbara H Partee, AGB ter Meulen et RE Wall. *Mathematical Methods in Linguistics*. Kluwer Academic Publishers, 1990.

[57] JM Rodrigues, B Trombert-Paviot, R Baud, J Wagner et F Meusnier-Carriot. Galen-in-use : using artificial intelligence terminology tools to improve the linguistic coherence of a national coding system for surgical procedures. In *Medinfo*, pages 623–627, 1998.

[58] Stefan Schulz, Anand Kumar et Thomas Bittner. Biomedical ontologies : What part-of is and isn't. *Journal of Biomedical Informatics*, 39(3) :350–361, 2006.

[59] Patrick Séguéla et Nathalie Aussenac. Un modèle de base de connaissances terminologiques. In *Terminologie et Intelligence Artificielle (TIA)*, pages 47–68, Toulouse, 1997.

[60] Barry Smith, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L Rector et Cornelius Rosse. Relations in biomedical ontologies. *Genome Biology*, 6 :R46, 2005.

[61] Barry Smith, Jennifer Williams et Steffen Schulze-Kremer. The ontology of the gene ontology. In *AMIA 2003*, pages 609–613, 2003.

[62] Kent Spackman et Keith Campbell. Compositional concept representation using SNOMED : Towards further convergence of clinical terminologies. In *Journal of American Medical Informatics Association (JAMIA)*, pages 740–744, 1998.

[63] Yannick Toussaint et Arnaud Simon. Building and interpreting term dependencies using association rules extracted from Galois lattices. In *Recherche d'Information Assistée par Ordinateur (RIAO)*, pages 1686–1692, Paris, April 2000.

[64] C.J. Wroe, R.D. Stevens, C.A. Goble et M. Ashburner. A methodology to migrate the Gene Ontology to a description logic environment using DAML+OIL. In *Pacific Symposium on Biocomputing (PSB)*, pages 624–636, 2003.

[65] Iwei Yeh, Peter D. Karp, Natalya F. Noy et Russ B. Altman. Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). *Bioinformatics*, 19(2) :241–248, 2003.