TALN 2009 – Session posters , Senlis, 24–26 juin 2009

Profilage sémantique endogène des relations de synonymie au sein de Gene Ontology

Thierry Hamon[1], Natalia Grabar[2]
(1)LIPN – UMR 7030, Université Paris 13 – CNRS, 99 av. J-B Clément, F-93430 Villetaneuse, France
(2)Centre de Recherche des Cordeliers, Université Pierre et Marie Curie - Paris6, UMR_S 872, Paris, F-75006 France ; Université Paris Descartes, UMR_S 872, Paris, F-75006 France ; INSERM, U872, Paris, F-75006 France ; HEGP AP-HP

Résumé. Le calcul de la similarité sémantique entre les termes repose sur l'existence et l'utilisation de ressources sémantiques. Cependant de telles ressources, qui proposent des équivalences entre entités, souvent des relations de synonymie, doivent elles-mêmes être d'abord analysées afin de définir des zones de fiabilité où la similarité sémantique est plus forte. Nous proposons une méthode d'acquisition de synonymes élémentaires grâce à l'exploitation des terminologies structurées au travers l'analyse de la structure syntaxique des termes complexes et de leur compositionnalité. Les synonymes acquis sont ensuite profilés grâce aux indicateurs endogènes inférés automatiquement à partir de ces mêmes terminologies (d'autres types de relations, inclusions lexicales, productivité, forme des composantes connexes). Dans le domaine biomédical, il existe de nombreuses terminologies structurées qui peuvent être exploitées pour la constitution de ressources sémantiques. Le travail présenté ici exploite une de ces terminologies, Gene Ontology.

Abstract. Computing the semantic similarity between terms relies on existence and usage of semantic resources. However, these resources, often composed of equivalent units, or synonyms, must be first analyzed and weighted in order to define within them the reliability zones where the semantic similarity shows to be stronger. We propose a method for acquisition of elementary synonyms which is based on exploitation of structured terminologies, analysis of syntactic structure of complex (multi-unit) terms and their compositionality. The acquired synonyms are then profiled thanks to endogenous indicators (other types of relations, lexical inclusions, productivity, form of connected components), which are automatically inferred within the same terminologies. In the biomedical area, several structured terminologies have been built and can be exploited for the construction of semantic resources. The work we present in this paper, is applied to terms of one of these terminologies, *i.e.* the Gene Ontology.

Mots-clés : Terminologie, distance sémantique, relations sémantiques, synonymie.

Keywords: Terminology, semantic distance, semantic relations, synonymy.

1 Introduction

Il est important de pouvoir détecter et calculer la similarité sémantique entre les termes, comme dans les exemples suivants : *acetone anabolism* et *acetone biosynthesis* ; *replication of mitochondrial DNA* et *mtDNA replication*, et *acetone catabolism* et *acetone breakdown*. En effet, différentes applications du traitement automatique de langue (TAL) – comme la recherche d'information, expansion de requêtes, extraction de connaissances, appariement de terminologies, fusion de bases de données – en sont demandeuses. Les lexiques de synonymes et de variantes morphologiques peuvent être utilisés pour le calcul de la similarité sémantique. En fonction des langues et des domaines, ces lexiques sont plus ou moins complets. La morphologie de la langue générale est la mieux décrite au travers des bases de données dédiées comme Celex (Burnage, 1990) ou MorTAL (Hathout *et al.*, 2001). Quant à la langue biomédicale, à laquelle nous nous intéressons dans ce travail, elle dispose surtout de ressources en anglais décrites dans l'UMLS Specialized Lexicon (NLM, 2007), mais également de ressources similaires dans d'autres langues (Schulz *et al.*, 1999; Zweigenbaum *et al.*, 2003), dont le français. Par contre, très peu est fait pour la constitution de ressources de synonymie : WordNet (Fellbaum, 1998) propose des synonymes pour l'anglais général, mais les ressources correspondantes pour d'autres langues ne sont pas librement disponibles. Par ailleurs, WordNet est insuffisant pour le traitement de la langue biomédicale (Bodenreider *et al.*, 2003) et l'initiative de son adaptation à ce domaine (Smith & Fellbaum, 2004) semble avoir été abandonnée. Nous pouvons donc constater un vrai manque de ressources de synonymie dans le domaine biomédical. Par contre, ce domaine propose plusieurs terminologies (*i.e.*, Gene Ontology (Gene Ontology Consortium, 2001), Snomed (Côté *et al.*, 1997) ou MeSH (NLM, 2001)) qui recensent des termes complexes et souvent leurs synonymes. L'utilisation des termes biomédicaux complexes ne semble pas être facile ni généralisable par les outils du TAL (Poprat *et al.*, 2008), tandis que les ressources lexicales, proposant des entités lexicales plus courtes et maniables, assurent cette fonction avec plus de naturel.

Dans notre travail, nous proposons d'utiliser les terminologies existantes pour inférer et ensuite profiler un lexique de synonymes élémentaires du domaine biomédical. Nous verrons que de tels synonymes sont en effet souvent « cachés » au sein de termes complexes. Notre méthode exploite essentiellement la compositionnalité des termes complexes et d'autres informations sémantiques endogènes accessibles au travers les mêmes terminologies.

2 Matériel et méthode

Nous travaillons avec Gene Ontology (GO) (Gene Ontology Consortium, 2001), qui propose un vocabulaire structuré et contrôlé pour la description des rôles de gènes et de leurs produits dans différents organismes. Les termes GO appartiennent à trois hiérarchies : processus biologiques, fonctions moléculaires et composants cellulaires. Les termes sont structurés avec quatre types de relations : subsomption `is-a`, méronymie `part-of`, synonymie et une relation transversale `regulates`. Nous avons utilisé la version de Gene Ontology téléchargée le 19 février 2008. Cette version contient 26 057 concepts et 79 994 libellés ou termes, ce qui correspond à une moyenne de 3,07 termes synonymes par concept. Ces termes sont reliés entre eux avec 433 294 relations `is-a`, 34 032 `part-of` et 460 780 relations de synonymie. Ce sont ces paires de termes que nous utilisons dans notre travail.

FIG. 1 – Analyse syntaxique de termes synonymes : *replication of mitochondrial DNA* et *mtDNA replication*.

Les termes GO sont souvent formés sur un schéma compositionnel (Verspoor *et al.*, 2003; Mungall, 2004; Ogren *et al.*, 2005). Selon la notion de compositionnalité, le sens d'une expression complexe est déterminé par sa structure syntaxique, le sens de ses composants et la fonction de composition (Partee, 1984). Par exemple, le concept GO :0009073 contient les synonymes qui dévoilent leur structure compositionnelle au travers la substitution d'un de leurs composants (composants soulignés dans les exemples) :

aromatic amino acid family biosynthesic process, *aromatic amino acid family biosynthesis*, *aromatic amino acid family anabolism*
aromatic amino acid family formation, *aromatic amino acid family synthesis*

Nous exploitons ce principe pour l'induction de synonymes élémentaires. Il permet ainsi d'induire la série suivante de synonymes élémentaires :

biosynthesic process, *biosynthesis*, *anabolism*, *formation*, *synthesis*

Cette approche exploite le travail précédent (Hamon & Nazarenko, 2001), où les auteurs avaient pour objectif de détecter des relations de synonymie entre termes complexes étant donné des relations de synonymie entre les termes simples. Dans le présent travail, la fonction inverse de la compositionnalité est exploitée : à partir des termes complexes, nous induisons des termes simples. Nous avons décrit l'implémentation et la généralisation de cette approche dans (Hamon & Grabar, 2008) et mentionnons ici, pour la clarté de l'exposé, uniquement ses points principaux (sec. 2.1). Nous décrivons ensuite les critères pour le profilage des relations de synonymie (sec. 2.2).

2.1 Acquisition de relations sémantiques élémentaires

La première étape de la méthodologie mise en oeuvre consiste à effectuer le pré-traitement des termes. Lors du pré-traitement, nous effectuons la segmentation des termes de la terminologie en mots, l'étiquetage morpho-syntaxique des termes et leur lemmatisation avec TreeTagger (Schmid, 1994). Nous effectuons ensuite une analyse syntaxique des termes et calculons les dépendances syntaxiques avec l'outil YaTeA[1]. Tous ces traitements sont réalisés au travers la plateforme `Ogmios`[2] adaptée au traitement de grands volumes de données et au domaine de la biomédecine. À l'issu de cette étape, chaque terme est considéré comme un arbre syntaxique binaire (comme sur la fig. 1) composé de deux composants : composant principal (*replication*) et dépendance (*mitochondrial DNA* et *mtDNA*).

La deuxième étape consiste en application des règles de compositionnalité. Ainsi, deux termes complexes synonymes $A\ rel\ B$ et $A\ rel\ B'$ avec le sens $\mathcal{M}$ ont la représentation sémantique suivante (le sens des termes complexes est fonction des sens de leurs composants) :

$$\mathcal{M}(A\ rel\ B) = f(\mathcal{M}(A), \mathcal{M}(B), \mathcal{M}(rel))$$

[1] http ://search.cpan.org/∼thhamon/Lingua-YaTeA/
[2] http ://search.cpan.org/∼thhamon/Alvis-NLPPlatform/

$$\mathcal{M}(A\ rel\ B') = f(\mathcal{M}(A), \mathcal{M}(B'), \mathcal{M}(rel))$$

À partir de ce principe, pour ces deux termes $A\ rel\ B$ et $A\ rel\ B'$, si un des composants (dans cet exemple, le composant principal A) est identique, alors les composants de dépendance (B et B') sont synonymes. Cette règle permet d'induire la paire de synonymes élémentaires {*mitochondrial DNA*, *mtDNA*} à partir des synonymes complexes *replication of mitochondrial DNA* et *mtDNA replication*. La variation est aussi acceptée sur le composant principal (A et A') dans une paire $A\ rel\ B$ et $A'\ rel\ B$ (comme dans *acetone catabolism* et *acetone breakdown*), de même que sur les deux composants à la fois, comme dans $A\ rel\ B$ et $A'\ rel\ B'$: *nicotinamide adenine dinucleotide catabolism* et *NAD breakdown*, où l'un des composants (par exemple, {*catabolism*, *breakdown*}) correspond à des synonymes déjà connus.

Cette méthode n'est pas spécifique à un type de relations. Elle fonctionne à un niveau assez abstrait de la langue – grâce à la représentation syntaxique – et est applicable à d'autres types de relations. Ainsi, avec le changement du type de la relation entre les termes complexes, le type de la relation entre leurs composants change de la même manière. Par exemple, si nous avons en entrée des termes complexes en relation `part-of` et si les règles de compositionnalité sont applicables, les termes élémentaires inférés auront la relation `part-of` entre eux. Ainsi, à partir des deux termes de processus biologiques *cerebral cortex development* GO :0021987 et *cerebral cortex regionalization* GO :0021796, nous inférons la relation élémentaire `part-of` entre leurs composants *development* et *regionalization*. Il en va de même pour les relations `is-a` : à partir de termes *cell activation* GO :0001775 et *astrocyte activation* GO :0048143 en relation `is-a`, nous pouvons établir la relation élémentaire `is-a` entre *cell* et *astrocyte*.

2.2 Profilage de relations sémantiques

Le profilage des relations sémantiques poursuit plusieurs objectifs : (1) faciliter la validation des relations de synonymie inférées, (2) pondérer ces relations pour garantir une meilleure spécificité lors de leur exploitation par un outil automatique, (3) respecter la nature contextuelle des relations sémantiques (Cruse, 1986). Pour profiler les relations de synonymie, nous combinons les indicateurs générés automatiquement et de manière endogène à partir de GO :

1. *Relations élémentaires* `is-a` acquises selon la méthode décrite en sec. 2.1 ;
2. *Relations élémentaires* `part-of` acquises selon cette même méthode décrite en sec. 2.1 ;
3. *Inclusion lexicale.* Les termes en relation de synonymie sont contrôlés pour l'inclusion lexicale (Bodenreider *et al.*, 2001). Si le test est positif, comme dans {*binding*, *DNA binding*}, les deux termes peuvent alors être en relation hiérarchique : la subsomption lexicale marque souvent la subsomption hiérarchique (Kleiber & Tamba, 1990) ;
4. *Productivité des relations de synonymie au sein de GO.* La productivité correspond au nombre de paires de synonymes originaux à partir desquelles une relation élémentaire est inférée. Par exemple, {*catabolism*, *breakdown*} est inféré à partir de six paires originales et {*nicotinamide adenine dinucleotide*, *NAD*} à partir de deux, tandis que {*anabolism*, *biosynthesis*} est inféré à partir de 15 paires. Ainsi, du point de vue de productivité de ces paires, {*anabolism*, *biosynthesis*} apparaît le plus fiable.
5. *Observation des paires de synonymes au sein des composantes connexes.* Les paires de synonymes élémentaires sont fusionnées et observées au sein des composantes connexes. La génération des composantes connexes est faite en exploitant les relations de ces paires de synonymes. Par exemple, à partir de deux relations de synonymes {*anabolism*, *biosynthesis*} et {*anabolism*, *synthesis*}, nous générons une composante connexe composée

de trois noeuds : *anabolism*, *biosynthesis* et *synthesis*. Le travail avec les composantes connexes de synonymes, plutôt qu'avec les paires de synonymes, nous permet de :

- situer les paires de synonymes dans un contexte plus global, où la synonymie est portée au niveau des familles de synonymes ;
- observer la forme des composantes connexes, qui peut être révélatrice de la force des relations de synonymie au sein des composantes connexes et donc de leur fiabilité ;
- explimer la force de synonymie de manière numérique, par exemple en calculant la densité des composantes connexes. Pour calculer la densité, nous prenons en compte le nombre effectif d'arrêtes $Nb - arrêtes$ et le nombre d'arrêtes que présenterait une clique (une composante connexe complète où tous les noeuds sont reliés entre eux) $Nb - arrêtes - clique$. La densité est calculée comme suit : $D = \frac{Nb-arrêtes}{Nb-arrêtes-clique}$ Selon cette formule, plus il y a de relations au sein d'une composante connexe, plus cette composante connexe est forte et plus la valeur de la densité D est proche de 1. Lorsque D est égal à 1, il s'agit d'une clique. Dans le cadre de profilage des relations de synonymie, plus la valeur de D est proche de 1 plus ces relations sont fortes et fiables.

La pondération des relations de synonymie est basée sur les principes suivants :

- Lorsqu'une relation de synonymie cooccurre avec un des trois premiers indicateurs et/ou lorsque sa productivité est faible cette relation s'en trouve affaiblie ;
- Lorsqu'une relation de synonymie apparaît toute seule (ne cooccurre pas avec un des trois premiers indicateurs) et elle montre une productivité élevée, et la composante connexe correspondante montre une densité élevée, alors il est fort probable qu'il s'agisse d'une relation de synonymie correcte.

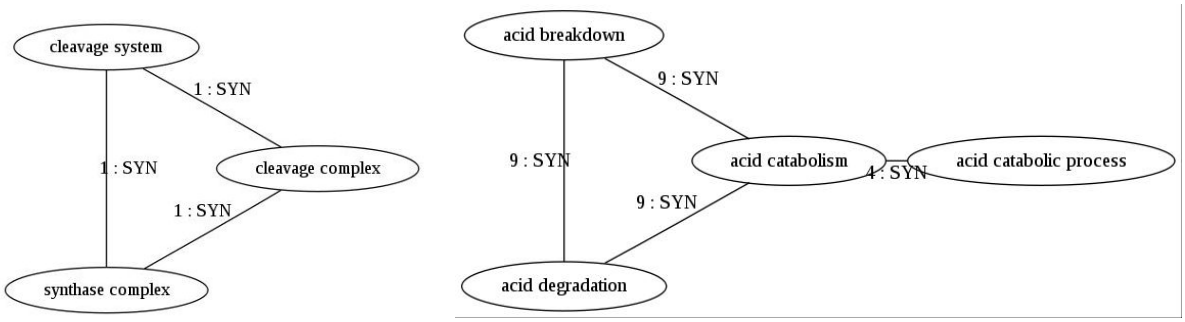
3 Résultats et discussion

3.1 Acquisition de relations sémantiques élémentaires

79 994 termes GO ont été analysés grâce à la plateforme Ogmios. Les règles de composition (sec. 2.1) ont été appliquées et ont permis d'induire 9 085 relations sémantiques élémentaires parmi lesquelles 3 031 relations de synonymie, 5 095 relations `is-a` et 1 540 `part-of`. L'inclusion lexicale produit 1 074 relations. Parmi les 3 031 relations de synonymie, 18 % (n=540) sont déjà présents dans la terminologie Gene Ontology. Les relations qui restent (n=2 491) correspondent à des relations « nouvelles ». Les 3 031 relations de synonymes élémentaires ont été groupées en 1 019 composantes connexes (CC), qui sont des ensembles de synonymes liés entre eux (fig. 2). Sur cette figure, nous pouvons observer une clique, où tous les noeuds sont reliés entre eux (fig. 2(a)), et une composante connexe (fig. 2(b)). Les arrêtes sont libellées par le type de la relation SYN et par la productivité.

3.2 Profilage des relations sémantiques

Pour le profilage des relations de synonymie, nous utilisons les indicateurs présentés dans la section 2.2 : relations élémentaires `is-a` et `part-of`, inclusions lexicales, productivité des relations de synonymie et la densité des composantes connexes. Nous allons appliquer ces indicateurs afin de décrire et profiler les composantes connexes et les relations de synonymie.



(a) Clique de synonymes (b) Composante connexe de synonymes

FIG. 2 – Composantes connexes formées de relations élémentaires de synonymie.

La figure 2(a) contient trois synonymes élémentaires (*cleavage system*, *cleavage complex*, et *synthase complex*) inférés à partir du concept GO :0004375. Cette composante contient trois synonymes qui sont tous reliés entre eux : il s'agit d'une clique. La densité de cette composante connexe est égale à 1. Des relations aussi denses entre les synonymes montrent que la force est grande dans cette famille. Par ailleurs, les relations de synonymie sont acquises à partir d'une paire de synonymes originaux de GO : la productivité est égale à 1. La combinaison de ces indicateurs, et surtout la forme de clique, montrent que ces relations de synonymie sont plutôt fiables.

En ce qui concerne la composante connexe présentée en 2(b), elle reproduit en grande partie le schéma du graphe 2(a). Elle se compose de quatre paires de synonymes élémentaires : {*acid degradation*, *acid breakdown*}, {*acid degradation*, *acid catabolism*}, {*acid breakdown*, *acid catabolism*} et {*acid catabolism*, *acid catabolic process*}. Les trois premières relations de synonymie forment une clique : la densité est maximale et la productivité est assez élevée (9 paires de synonymes originaux). Par contre, la paire de synonymes {*acid catabolism*, *acid catabolic process*} reste à part. Elle est acquise à partir de quatre paires de synonymes (notamment, *glutamic acid catabolism* et *fatty acid catabolism*), ce qui correspond à une productivité également assez élevée. Comme cette relation reste assez excentrée dans la composante connexe de la figure 2(b) et comme sa productivité est moins élevée que celle des autres relations, elle peut être considérée comme le point faible de cette composante connexe.

Le graphe de la figure 3 montre la coocurrence de la synonymie avec d'autres relations sémantiques : *INCL* marque la relation hiérarchique induite avec l'inclusion lexicale, *HIER* marque la relation `is-a` induite par la composition, *PAR* marque la relation `part-of` induite par la composition. Ainsi, les paires de synonymes élémentaires {*genome*, *chromosome*} et {*DNA*, *chromosome*} sont induites à partir des relations de synonymie et de méronymie. Quant à la paire {*chromosome*, *interphrase chromosome*}, elle est induite à partir des relations de synonymie et d'hypéronymie mais également calculée au travers de l'inclusion lexicale. Il s'agit donc d'une relation hiérarchique potentielle. Avec l'analyse de la forme de cette composante connexe, nous remarquons qu'elle a la forme d'une étoile, ce qui correspond à une structure beaucoup moins fiable qu'une clique (comme sur la figure 2(a)). Au sein de cette composante connexe, nous pouvons identifier trois zones avec *DNA* comme pivot :

1. composants de la cellule : *chromosome*, *interphase chromosome*, *genome* ;
2. processus biologiques : *P-element*, *cut-and-paste* ;
3. une relation de synonymie {*DNA*, *polydeoxyribonucleotide*}.

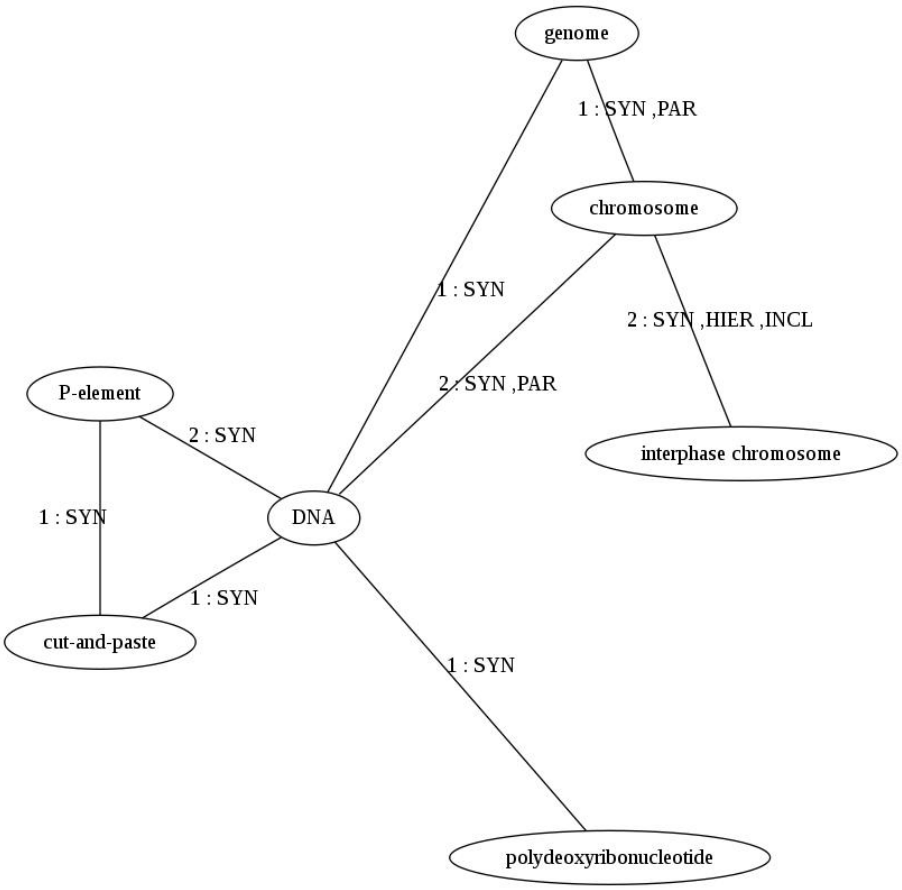


FIG. 3 – Relations élémentaires de synonymie qui cooccurrent avec d'autres types de relations.

Les deux premières zones forment des cliques : la zone 1 est une clique et la zone 2 contient une clique. La zone 3 est également une clique mais minimale (avec deux noeuds seulement). L'étude de l'occurrence des indicateurs avec les relations de synonymie nous aide à mieux appréhender la nature des relations au sein de cette composante connexe, ou même au sein de ses cliques, et à les profiler. Regardons de plus près la zone 1 composée de *chromosome*, *interphase chromosome*, *genome* et *DNA*. Comme nous l'avons noté, les relations {*DNA*, *chromosome*} et {*genome*, *chromosome*} sont également induites à partir des relations de méronymie et leur sémantique est donc ambiguë, bien qu'elles fassent partie d'une clique. Quant à la relation {*chromosome*, *interphase chromosome*}, qui est excentrée dans cette zone, elle est ambiguë avec la relation d'hyperonymie au travers de deux indicateurs (inclusion lexicale et relation `is-a` élémentaire). La zone 1 de cette composante connexe montre donc des points de faiblesse malgré le fait qu'elle comporte une clique. Nous pouvons donc voir que l'analyse d'une composante connexe devrait combiner les indicateurs sémantiques, la productivité des relations et la forme de cette composante. Ceci permet de détecter des zones plus fortement connexes que d'autres et éventuellement conduire à un partitionement du graphe afin de former des sous-graphes sémantiquement plus cohérents et fiables.

Étant donné ces considérations, sur la figure 4 nous pouvons distinguer deux zones fortement connexes : (1) *regulation*, *control*, *modulation*, *regulator* ; (2) *cell cycle control*, *cell cycle regulation*, *cell cycle regulator*, *cell cycle modulation*. Ces deux zones peuvent justement correspondre aux partitions principales de la composante connexe. De plus, l'exploitation de la

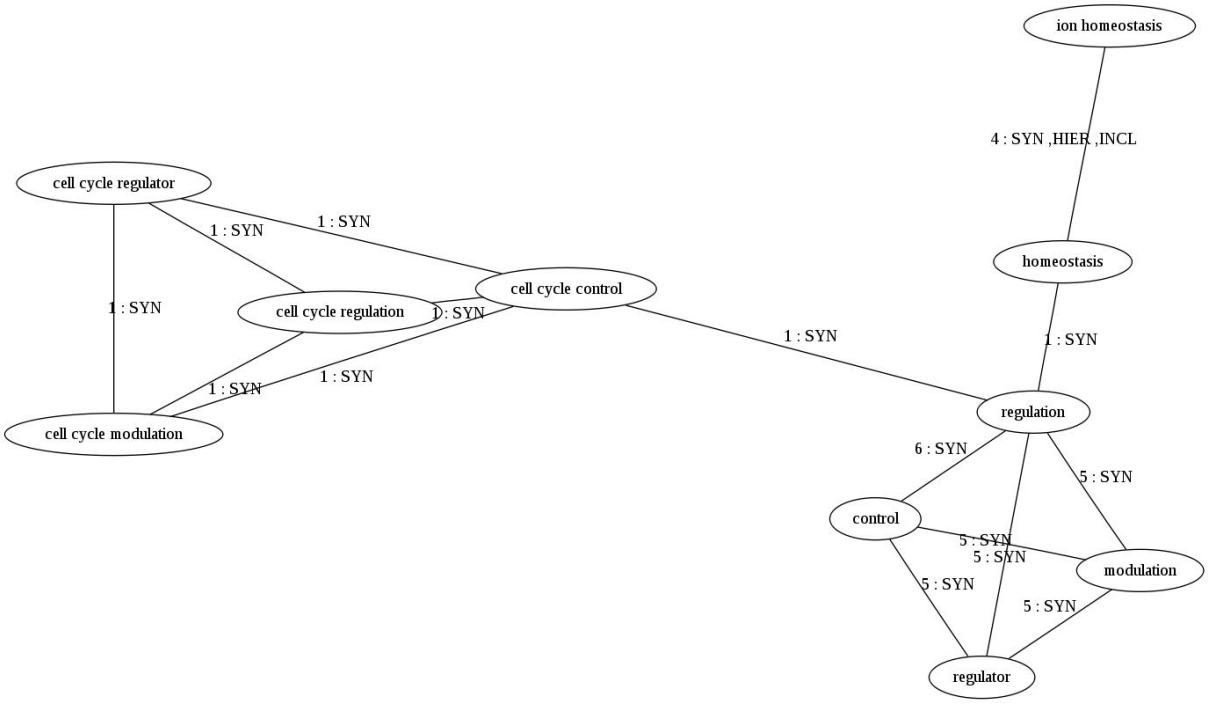


FIG. 4 – Identification de zones fortement et exploitation de la pondération des arêtes.

productivité des relations élémentaires au sein de la zone 1, qui est assez importante (entre 5 et 6), permet de conforter cette hypothèse. Le profilage des relations élémentaires de synonymie selon cette méthodologie permet donc de renforcer les relations au sein de la zone 1, de les renforcer un peu au sein de la zone 2, et d'affaiblir la validité de la relation {*cell cycle control*, *regulation*}. Notons que nous n'avons pas analysé la zone composée des noeuds *regulation*, *homeostasis* et *ion homeostasis*.

Une évaluation manuelle des relations élémentaires de synonymie a été effectuée par un informaticien avec, pour critère principal, le contexte original d'apparition des synonymes élémentaires. Cette évaluation a montré une précision de 93,1 %, avec 5,4 % de relations rejetées et 1,5 % relations toujours sous question. Il est clair que la perception de ces synonymes ne sera pas la même lorsque la validation est effectuée par un expert du domaine de biologie. Nous pensons que l'évaluation actuelle des relations de synonymie montre un profilage beaucoup plus fin et fournit des indicateurs de validité plus fiables à présenter à un expert du domaine. De même, lorsque ces indicateurs sont formalisés et chiffrés, ils peuvent être utilisés par des outils automatiques.

4 Conclusion and Perspectives

Nous avons proposé une méthode pour l'inférence de relations de synonymie élémentaires à partir de terminologies structurées et pour leur profilage en vue d'une pondération. La méthode exploite la compositionnalité et les dépendances syntaxiques des termes complexes. Les indicateurs de profilage sont générés automatiquement et de manière endogène à partir de la même terminologie. Le travail a été effectué sur les termes de *Gene Ontology*, mais peut être appliqué à d'autres terminologies structurées et d'autres domaines.

Si les relations et les indicateurs sont générés automatiquement, le profilage est fait actuellement de manière empirique : il ne correspond pas à une pondération mais plutôt à une décription. Il sera ainsi intéressant de modéliser cette méthodologie et de l'implémenter pour la pondération des relations de synonymie. Le calcul de la pondération ou de la probabilité que l'on pourra associer ainsi à une relation de synonymie peut par exemple être effectuée avec une approche par réseaux bayesiens (Spirtes *et al.*, 1993; Pourret *et al.*, 2008). Cette perspective demande une collaboration étroite entre les chercheurs en informatique et en biologie. Ainsi, lorsque, au sein d'une composante connexe, certaines arrêtes montrent un poids faible, ces arrêtes peuvent être supprimées et le graphe décomposé en sous-graphes plus petits et possiblement plus fiables. Dans le travail présenté ici, nous avons visé la pondération des relations de synonymie par les indicateurs de fiabilité. Notons que le même travail peut être effectué vis-à-vis d'autres types de relations (par exemple, profilage des relations hiérarchiques par les indicateurs de fiabilité). De cette manière, une validation mutuelle des relations sémantiques pourrait être obtenue.

Les relations inférées pourront être généralisées au travers des corpus et servir à l'enrichissement et adaptation d'une terminologie (Hamon *et al.*, 1998; Hole & Srinivasan, 2000). Du point de vue d'une perspective ontologique, cette méthode est un pas vers la vérification de la consistance d'une terminologie (Mungall, 2004). Elle peut aussi être utilisée pour la transformation d'une terminologie pré-coordonnée en une terminologie post-coordonnée.

Références

BODENREIDER O., BURGUN A. & MITCHELL J. A. (2003). Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases : a feasibility study. In *Medical Informatics in Europe (MIE)*, p. 379–384.

BODENREIDER O., BURGUN A. & RINDFLESCH T. C. (2001). Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. In URI INIST CNRS, Ed., *Terminologie et Intelligence artificielle (TIA)*, p. 11–21, Nancy.

BURNAGE G. (1990). *CELEX - A Guide for Users*. University of Nijmegen : Centre for Lexical Information.

CÔTÉ R. A., BROCHU L. & CABANA L. (1997). *SNOMED Internationale – Répertoire d'anatomie pathologique*. Secrétariat francophone international de nomenclature médicale, Sherbrooke, Québec.

CRUSE D. A. (1986). *Lexical Semantics*. Cambridge : Cambridge University Press.

FELLBAUM C. (1998). A semantic network of english : the mother of all WordNets. *Computers and Humanities. EuroWordNet : a multilingual database with lexical semantic network*, **32**(2-3), 209–220.

GENE ONTOLOGY CONSORTIUM (2001). Creating the Gene Ontology resource : design and implementation. *Genome Research*, **11**, 1425–1433.

HAMON T. & GRABAR N. (2008). How can the term compositionality be useful for acquiring elementary semantic relations ? In B. NORDSTROM & A. RANTA, Eds., *GoTAL*, number 5221 in LNAI, p. 809–814, Gothenburg, Sweden : Springer.

HAMON T. & NAZARENKO A. (2001). Detection of synonymy links between terms : experiment and results. In *Recent Advances in Computational Terminology*, p. 185–208. John Benjamins.

HAMON T., NAZARENKO A. & GROS C. (1998). A step towards the detection of semantic variants of terms in technical documents. In *International Conference on Computational Linguistics (COLING-ACL'98)*, p. 498–504, Université de Montréal, Montréal, Quebec, Canada.

HATHOUT N., NAMER F. & DAL G. (2001). An experimental constructional database : the MorTAL project. In P. BOUCHER, Ed., *Morphology book*. Cambridge, MA : Cascadilla Press.

HOLE W. & SRINIVASAN S. (2000). Discovering missed synonymy in a large concept-oriented metathesaurus. In *AMIA 2000*, p. 354–8.

KLEIBER G. & TAMBA I. (1990). L'hyperonymie revisitée : inclusion et hiérarchie. *Langages*, **98**, 7–32.

MUNGALL C. (2004). Obol : integrating language and meaning in bio-ontologies. *Comparative and Functional Genomics*, **5**(6-7), 509–520.

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

NLM (2007). *UMLS Knowledge Sources Manual*. National Library of Medicine, Bethesda, Maryland. www.nlm.nih.gov/research/umls/.

OGREN P., COHEN K. & HUNTER L. (2005). Implications of compositionality in the Gene Ontology for its curation and usage. In *Pacific Symposium of Biocomputing*, p. 174–185.

PARTEE B. H. (1984). *Compositionality*, In *Varieties of formal semantics*. F Landman and F Veltman.

POPRAT M., BEISSWANGER E. & HAHN U. (2008). Building a biowordnet using wordnet data structures and wordnet's software infrastructure - a failure story. In *ACL 2008 workshop "Software Engineering, Testing, and Quality Assurance for Natural Language Processing"*, p. 31–9.

POURRET O., NAIM P. & MARCOT B. (2008). *Bayesian Networks : A Practical Guide to Applications*. Chichester, UK : Wiley.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, p. 44–49, Manchester, UK.

SCHULZ S., ROMACKER M., FRANZ P., ZAISS A., KLAR R. & HAHN U. (1999). Towards a multilingual morpheme thesaurus for medical free-text retrieval. In *Medical Informatics in Europe (MIE)*.

SMITH B. & FELLBAUM C. (2004). Medical wordnet : a new methodology for the construction and validation of information. In *Proc of 20th CoLing*, p. 371–382, Geneva, Switzerland.

SPIRTES P., GLYMOUR C. & SCHEINES R. (1993). *Causation, Prediction, and Search*. New York : Springer-Verlag.

VERSPOOR C. M., JOSLYN C. & PAPCUN G. J. (2003). The gene ontology as a source of lexical semantic knowledge for a biological natural language processing application. In *SIGIR workshop on Text Analysis and Search for Bioinformatics*, p. 51–56.

ZWEIGENBAUM P., BAUD R., BURGUN A., NAMER F., JARROUSSE E., GRABAR N., RUCH P., DUFF F. L., THIRION B. & DARMONI S. (2003). Towards a Unified Medical Lexicon for French. In *Medical Informatics in Europe (MIE)*.