La relation de synonymie en génomique

Davy Weissenbacher

LIPN – Université de Paris 13
99, avenue Jean-Baptiste Clément
93 430 Villetaneuse
dw@*lipn.univ-paris13.fr*

Résumé – Abstract

L'accès au contenu des textes de génomique est aujourd'hui un enjeu important. Cela suppose au départ d'identifier les noms d'entités biologiques comme les gènes ou les protéines. Se pose alors la question de la variation de ces noms. Cette question revêt une importance particulière en génomique où les noms de gènes sont soumis à de nombreuses variations, notamment la synonymie. A partir d'une étude de corpus montrant que la synonymie est une relation stable et linguistiquement marquée, cet article propose une modélisation de la synonymie et une méthode d'extraction spécifiquement adaptée à cette relation. Au vu de nos premières expériences, cette méthode semble plus prometteuse que les approches génériques utilisées pour l'extraction de cette relation.

The access to textual content in genomics is now recognized as an important issue. One of the first steps is the recognition of biological entity names such as gene or protein names. It has often been observed that entity names may vary in texts but this phenomenon is especially common in genomics. Besides a gene canonical name, one can find various abbreviation forms, typographic variants and synonyms. Stemming in a corpus analysis, this paper argues that synonymy in genomic texts is a stable and linguistically marked relation. This paper presents a method for extracting couples of synonymous gene or protein names. From a preliminary experiment, this method seems more promising than generic approaches that are exploited to extract synonymy relations.

Keywords – Mots Clés

Extraction d'information, synonymie, entités nommées, génomique

Information extraction, synonymy, named entities, genomics

1 Introduction

La génomique est un domaine pour lequel le manque d'outils permettant d'explorer efficacement le contenu des bases documentaires est critique. Les chercheurs ont besoin de retrouver rapidement un complément d'information fiable sur un objet biologique auquel ils s'intéressent. Des moteurs de recherche spécialisés existent, Medline [1] par exemple, mais leurs réponses même aux questions les plus ciblées, ramènent beaucoup trop de documents pour être consultés humainement dans un temps imposé. La qualité rédactionnelle des articles, leur homogénéité et le fait qu'ils relèvent d'un domaine spécialisé poussent à développer des systèmes d'extraction d'information spécialisés. C'est la piste qu'explorent différents groupes de recherche (Caderige, Genia, Biomint[2]).

Pour améliorer l'accès à l'information textuelle, le repérage des noms de gènes et de protéines, entre autres substances biologiques, est capital. Toutes les méthodes classiques d'extraction de ces *entités nommées*[3] (ENs) visant à les repérer en corpus se heurtent aux problèmes de la variation synonymique. Ce phénomène a une ampleur particulière en génomique : 1) un très grand nombre d'ENs sont pourvues de synonymes [4] (environ 40% des noms de gènes sont référencés dans Flybase[5] avec au moins un synonyme) ; 2) une même EN peut avoir beaucoup de synonymes (ex. le gène *Hcph* a au moins 6 synonymes). Dans ce domaine de la génomique, négliger ces phénomènes de variation ou les sous-estimer, comme cela a souvent été le cas, limite la portée des travaux effectués pour le repérage des ENs.

La présente étude tente d'analyser ces phénomènes de synonymie. Nous défendons l'idée selon laquelle, dans les corpus de génomique, cette relation est facile à repérer dès lors que l'on tient compte de ses propriétés spécifiques. Nous proposons une méthode d'extraction adaptée au repérage des liens de synonymie entre noms de gènes ou de protéines. Au vu de nos premières expériences, cette méthode semble plus prometteuse que les approches génériques utilisées pour la variation des entités nommées.

La section 2 de cet article présente l'état des travaux sur le repérage des relations de synonymie. L'analyse en corpus des relations de synonymie entre noms de gènes ou de protéines (section 3) nous conduit à proposer une approche originale pour extraire des couples de synonymes (section 4). Nous montrons dans la section 5 une expérience préliminaire qui permet de valider *a priori* l'approche proposée. La section 6 discute l'apport de notre approche et ses perspectives possibles.

---

1 URL : http://www.ncbi.nlm.nih.gov/pubmed/

2 URL : http://caderige.imag.fr/ ; http://www.biomint.org/ ; http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

3 Une *entité nommée* est un syntagme qui renvoie à un unique objet d'une réalité supposée (selon les contextes, il s'agit des noms propres de personnes, des dates, des URLs, etc.).

4 Nous considérons comme synonymes deux ENs qui désignent la même entité et qui peuvent être substituées l'une à l'autre sans modifier le sens de l'expression où elles figurent. Dans ce sens la synonymie couvre différents phénomènes : variantes typographiques, phénomènes de reformulations, *etc.*

5 Flybase : http://flybase.bio.indiana.edu/

2 Comment repérer des relations de synonymie dans les textes ?

Trois grandes approches ont été proposées pour repérer des liens de synonymie dans les corpus. La première, l'approche *distributionnelle* (voir (Lin, 1998) par exemple), est floue et bruitée. Elle rapproche bien deux unités mais sans typer le lien qu'elles entretiennent. Les regroupements produits ne peuvent être exploités sans être validés. Les expériences de (Agichtein, Yu, 2003) confirment cette analyse. La seconde approche est *contextuelle* : elle repose sur des patrons d'extraction pour repérer des liens de synonymie entre unités. (Pearson, 1998) s'appuie sur des marqueurs de relation comme *known as, called*. Approche prometteuse par sa simplicité de mise en œuvre et par sa précision, (Yu *et al*., 2002) l'exploite pour identifier des liens de synonymie entre des noms de gènes ou de protéines. Cependant les résultats obtenus (30% de précision seulement sur les résumés de Medline) sont décevants. A notre sens, les auteurs n'exploitent pas assez les marqueurs linguistiques : sont privilégiés les indices de bas niveau comme la ponctuation ou la casse. La troisième approche est *structurelle* (voir (Hole, Srinivasan, 2000) par exemple). Deux mots ou termes sont proches s'ils sont construits de manière similaire (structure et/ou constituants identiques ou voisins). Cette approche coûteuse en connaissances extérieures est une alternative intéressante lorsque l'approche contextuelle s'avère peu explo itable. En biologie, elle est utilisée pour identifier des abréviations très fréquentes dans les textes de génomique (Chang *et al.*, 2002).

Dans ce qui suit, nous montrons que l'analyse spécifique de l'expression de la synonymie dans les textes de génomiques permet de proposer une méthode d'acquisition ciblée, plus efficace, que ce qui a été proposé par (Yu *et al*., 2002) par exemple.

3 Etude de la relation de synonymie

Lorsqu'une phrase contient au moins deux ENs identifiées nous appelons *fragment* l'empan de texte dans lequel la relation de synonymie est exprimée et doit être recherchée[6]. Nous entendons par *amorce de la synonymie* le mot ou l'expression qui est utilisé pour signifier que des ENs sont synonymes (*also called* par exemple). *Un Fragment Introduisant la Synonymie* (FIS) est un fragment (au sens précédemment défini) qui contient une amorce de la synonymie et dont les deux ENs déterminées sont des synonymes ; un FIS peut contenir plusieurs relations de synonymies[7]. Le FIS est la partie de la phrase que l'on cherche à repérer et à analyser automatiquement. Enfin nous définissons la structure d'un FIS comme une séquence ordonnée d'unités normalisées, les indices.

---

[6] C'est l'ensemble des mots contigus situés entre la première EN et la dernière EN, si aucune ponctuation double n'est ouverte entre la première EN et la dernière EN sans être refermée avant la dernière EN (exemple : […]*factors termed IFN regulatory factors (IRF) and is also called IRF-8*.) ; ou entre la première EN et le signe de ponctuation double fermant, si une ponctuation double est ouverte entre la première EN et la dernière EN sans être refermée avant la dernière EN (exemple : […]*codes for threonine-tRNA ligase (tRNAThr ligase, formerly threonine-tRNA synthetase, EC 6.1.1.3) has previously*[…])

[7] Le fragment *IFN regulatory factors (IRF) and is also called IRF-8* est un FIS contenant trois synonymes : *IFN* regulatory factors, *IRF* et *IRF-8*.

Nous avons constitué un corpus de phrases pour étudier le comportement de la relation de synonymie. Nous avons fourni au moteur d'interrogation de Medline 12 couples de noms de gènes que nous savions être synonymes. Dans les résumés retournés, nous avons sélectionné manuellement toutes les phrases exprimant la synonymie autour d'un élément du couple. Le corpus de travail obtenu comporte 63 phrases exprimant la synonymie.

L'étude de notre corpus de travail montre tout d'abord que la synonymie est une relation fortement *redondante* : plusieurs articles expriment de manières différentes la synonymie entre deux noms de gène identiques (11 fois pour les noms de gènes Tal-1 et SCL). On constate également que l'expression de la synonymie est *locale* : nous n'avons pas rencontré de synonymie qui s'exprime contextuellement sur plusieurs phrases et toutes les informations nécessaires à l'expression de la relation de synonymie sont concentrées dans une partie de la phrase, que nous avons appelé le FIS. Enfin, on note que l'expression de la synonymie est relativement *stable* : nous n'avons distingué que 3 grands types de FIS subissant de petites variations.
Notre analyse révèle que les FIS sont structurés *i.e.* ce sont des séquences ordonnées de termes normalisés, les indices. Ces structures varient uniquement par l'ordre et la présence de ces indices. La figure n°1 représente les structures sans variations des 3 grands types de FIS. Nous avons logiquement choisi une modélisation sous forme d'arbres, modélisation capable de rendre compte de cette structure. Les nœuds des arbres représentent les indices de la synonymie et les feuilles les termes du fragment. Parmi les indices nous comptons : la ponctuation qui sert à séparer une partie du FIS du reste de la phrase, à introduire un commentaire dans le FIS, ou à distinguer les éléments d'une énumération (souvent terminées par les conjonctions *and* et *or*) ; les acronymes qui sont nombreux dans les FIS ; le type de l'entité qui est mentionné explicitement dans le contexte immédiat des entités synonymes (par exemple : *The AML1 gene[…]*) ; les amorces de la synonymie qui sont relativement figées (peu nombreuses et subissant peu de variations, exemple : *otherwise known as* et *also known as*) ; les éléments inconnus *i.e.* tout ensemble de termes contenu dans le FIS et qui ne sont pas des indices.

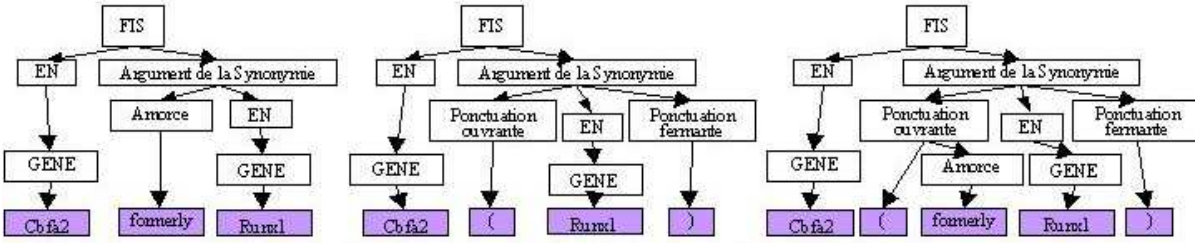

Figure1. Structures sans variations des 3 grand types de FIS

Nous défendons l'hypothèse que reconnaître la relation de synonymie c'est reconnaître la structure, avec ses variations, des FIS. La section suivante décrit la méthode suivie pour la valider.

4 Validation

Pour valider notre hypothèse, et de fait notre analyse, nous avons engendré 7 automates reconnaissant les 3 grands types de FIS et leurs variations (voir figure°2). Les états des automates sont les indices que nous supposons savoir reconnaître (à l'exception des éléments inconnus : chaque FIS contenant un élément inconnu sera ignoré).

Nous avons constitué un corpus de test, en projetant sur Medline l'amorce *formerly* de fiabilité moyenne[8]. Nous avons uniquement exigé des 106 phrases qui composent notre corpus de test qu'elles contiennent l'amorce *formerly* et ses variations, et au moins deux EN que nous avons annotées manuellement.

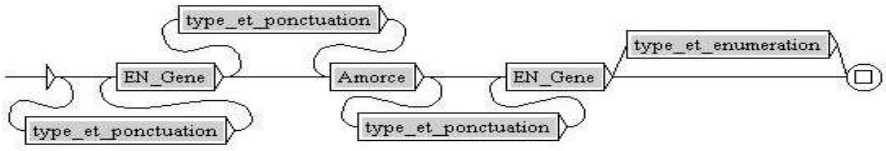

Figure2. Automate engendré par la 1ère structure de la figure°1 et ses variations

5 Résultats

Les résultats de la projection des automates résumés dans le tableau°1 ont été vérifiés par une relecture humaine du corpus de test. Un fragment reconnu par un automate et qui n'exprime pas une relation de synonymie est classé dans les fragments erronés. Tous les fragments exprimant une relation de synonymie non reconnue ou seulement partiellement reconnue par un ou plusieurs automates sont classés dans les fragments ignorés.

Fragments incorrects et refusés	9	Fragments corrects et ignorés	13
Fragments corrects et reconnus	69	Fragments corrects et partiellement reconnus	13
Fragments incorrectement reconnus	2		

Tableau1. Résultat de la projection des automates sur le corpus de test

La précision des automates est de 97,5%. Les deux fragments incorrectement reconnus sont des dépendances croisées. Aucun automate n'a été prévu pour reconnaître cette forme de FIS, cette dernière n'étant pas présente dans le corpus de travail. Le rappel, au score assez faible de 75%, demande à être nuancé. En premier lieu parce que nos automates sont incomplets. L'ajout de 5 variations imprévues dans les automates permet la reconnaissance de 7 fragments supplémentaires. Et en second lieu par la réhabilitation des éléments inconnus qui siègent au sein du fragment, car une analyse humaine leur accorde une identité[9]. Des études spécialisées aboutissant à l'identification automatique de ces éléments inconnus augmenteraient le rappel sans risquer de dégrader la précision de nos automates. Seules 4 fragments présentent de réels

---

[8] Certaines amorces, comme « / », introduisent beaucoup de bruits et d'autres très peu, comme « *also called* »

[9] 5 fragments sont entrecoupés par références bibliographiques exemple : *Heparinase III (E.C. 4.2.2.8), formerly heparinase I,* 5 fragments sont entrecoupés par la relation d'hyperonymie exemple : *Galectin-3 is an animal lectin, formerly named epsilon-binding,* 2 par le renommage exemple : *Spi 2.1, formerly Spi.1 has recently been redesignated Spin2a),* 2 par l'homologie exemple : *utrophin (the chromosome 6-encoded dystrophin homolog formerly known as dystrophin-related protein),* 2 par l'encodage exemple : *NusG is encoded by an E. coli gene, formerly called U and now called nusG*, et 1 par un type inconnu

obstacles théoriques. Le premier fragment conjugue une modalité à l'amorce de la synonymie. Le deuxième fragment, une dépendance croisée, ne fut que partiellement reconnu. Les deux derniers fragments juxtaposent 2 automates ce qui est une opération interdite dans notre représentation (ex. *XLH gene, referred to as PHEX, or formerly as PEX,[…]*).

6 Discussion et perspectives

En nous appuyant sur les caractéristiques de la relation de synonymie observées sur un petit corpus de génomique, nous avons proposé une représentation des fragments de phrases exprimant la synonymie sous la forme d'arbres. Nous mettons ainsi la structure synonymique en évidence au-delà de la séquence de mots et mots-clefs qui constituent le fragment de phrase exprimant la synonymie. Cette modélisation ne comporte qu'un très petit nombre de structures principales, chacune variant par l'ordre et la présence d'éléments seconda ires, les indices. Nous avons évalué la qualité et la couverture de cette description en calculant des automates à partir des structures arborescentes, ceux-là linéarisant ceux-ci. Nous avons montré que nos résultats de 97,5% de précision et de 75% de rappel sont intéressants et meilleurs pour les résumés de MedLine que les travaux antérieurs.

Nous poursuivons notre travail avec l'objectif d'exploiter le double intérêt de notre modélisation ; d'une part associer des informations (comme le type des synonymes ou le degré de fiabilité de l'amorce) à chaque structure en vue de calculer la fiabilité globale du fragment et d'autre part faciliter l'apprentissage des règles d'extraction par une approche modulaire : d'abord l'acquisition des indices (ENs, amorces, ac ronymes, etc.) par des apprentissages spécifiques à chacun d'eux, puis la reconnaissance de la structure du fragment.

Références

Agichtein E, Yu H. (2003), Extracting synonymous gene and protein terms from biological literature, *Bioinformatics*, vol. 19 Suppl;1.

Chang J.T., Schütze H., Altman R. (2002), Creating an online dictionary of abreviations from Medline, *Journal of the American Medical Informatics Association;pp. 612-620*

Hole W., Srinivasan S. (2000), Discovering missed synonyms in a large concept-oriented metathesaurus, Proceedings of *AMIA Symposium*, pp. 354-358.

Lin D. (1998), Automatic retrieval and clustering of similar words, Proceedings of *ACL'98*.

Pearson J (1998), *Terms in Context*, John Benjamins.

Yu H. Hatzivassiloglou V., Friedman C., Rzhetsky A., Wilbur J.(2002), Automatic Extraction of gene and protein synonyms from medline and journal articles, Proceedings of AMIA Symposium, pp. 413-423.