Recherche d'information et analyse bibliographique appliquées à la mise à jour automatique de Swiss-Prot

Imad Tbahriti[12] — Anne-Lise Veuthey[2] — Patrick Ruch[1] — Julien Gobeill[12]

[1] Service d'Informatique Médicale, Université de Genève

imad.tbahriti@medecine.unige.ch

[2] Groupe Swiss-Prot, Institut Swiss de Bioinformatique

Catégorie : chercheur

RÉSUMÉ.

But : Le but de cette étude est de découvrir de nouveaux articles scientifiques utiles pour la mise à jour de l'information dans la base de données de biologie moléculaire UniProtKB/Swiss-Prot. Notre hypothèse de base est qu'un article qui cite un autre article déjà référencé dans une entrée Swiss-Prot pour une protéine donnée est un bon candidat pour mettre à jour l'information de l'entrée de cette protéine dans la base.

Méthodes : La procédure expérimentale pour tester cette hypothèse est la suivante : dans chaque entrée UniProtKB/Swiss-Prot nous séparons l'ensemble des références bibliographiques connues (ERB) en deux ensembles : l'ensemble des références connues (ERC) et l'ensemble des références à découvrir (ERD) par notre système. Notre système va proposer un ensemble de références possibles (ERP). Nous évaluons la performance de deux différentes méthodes en comparant la précision de notre détecteur de nouveauté, c'est-à-dire en mesurant le rapport entre ERD et ERP. La première méthode, guidée par les références (GR) analyse les références bibliographiques d'un article donné pour prévoir son intérêt pour la mise à jour de UniProtKB/Swiss-Prot. Pour cette approche, nous utilisons un analyseur HTML de bibliographie afin d'identifier dans notre corpus les articles citant des articles contenus dans ERB. La deuxième méthode, guidée par la recherche documentaire (GD) utilise un moteur de recherche pour classer par ordre de pertinence un ensemble d'articles en fonction d'une requête contenant le nom de la protéine et ses synonymes. Pour cette approche, nous avons récupéré l'ensemble des champs MEDLINE; résumé, titre, termes MeSH (Medical Subject Headings ou Mots clés médicaux) et les noms chimiques correspondants à ces PMID afin de les indexer dans le moteur de recherche.

Résultats : On trouve une importante corrélation (0.74) entre les deux méthodes GR et GD. Toutefois des différences demeurent, en particulier, la précision des premiers documents retournés par méthode GD est sensiblement supérieure (0.46) à celle de la méthode GR

CORIA 2007, pp. 319–334
Saint-Étienne, 28–30 mars 2007

(0.31). La fusion des deux méthodes par combinaison linéaire, qui obtient un gain significatif en précision de +6.5% confirme la complémentarité des deux méthodes.

Conclusion : Nos résultats montrent qu'une approche basée sur une méthode bibliométrique utilisant des réseaux de citations, est complémentaire d'une approche basée sur la recherche d'information pour détecter les nouvelles connaissances utiles pour annoter la base de données UniProtKB/Swiss-Prot.

ABSTRACT.

Purpose: The goal of this study is to discover new articles valuable for updating the information in the UniProtKB/Swiss-Prot database. Our basic hypothesis is that an article that cites a PubMed reference (PMID) found in an entry in the Swiss-Prot database will be a good candidate for updating that specific protein entry. We want to verify this hypothesis and validate it by comparing this approach against and in combination with a search-based method.

Methods: To test our hypothesis we separated the known bibliographic references (ERB) from each UniProtKB/Swiss-Prot entry into two groups: the set of the known references (ERC) and the set of the references to discover (ERD). Our system will propose some candidate references (ERP) that cite the known references found in UniProtKB/Swiss-Prot entries (ERB), which will have to be evaluated. We tested two different methods to find the ERP useful for updates of the UniProtKB/Swiss-Prot records. For each candidate reference (ERP) proposed by our system by one of the two methods, we evaluated the effectiveness by comparing the precision, i.e. by measuring the relationship between ERD and ERP. The first method guided by the references (GR) analyzes the bibliographical references of a given article to predict its benefit for the update of UniProtKB/Swiss-Prot. For this approach, we analyzed the citations from all articles in our corpus in order to identify those which reference the articles contained in the ERB. The second method (GD) uses an information retrieval engine to rank a set of articles in terms of a query containing the protein name and its synonyms. For the second approach, we recovered from MEDLINE the abstract, title, MeSH terms and chemical names for each PMID as input for information retrieval engine

Results: We found a significant correlation (0.74) between the article ranking given by the information retrieval engine and the article ranking given by GR method. However, the precision at high ranks of the GR method (0.31) is lower than the GD method (0.46). The fusion of the two approaches by linear combination significantly improves the baseline (GD) by +6.5%. Thus, confirming that the two methods are complementary.

Conclusion: Our results show that an approach based on a bibliometric method using citation networks, is an informative and novel method to provide information appropriate for the updating of the UniProtKB/Swiss-Prot database.

MOTS-CLÉS : *protéomique, bibliothèque digitale, bibliométrie, fouille de données, recherche documentaire, intelligence artificielle, détection de nouveautés, citation.*

KEYWORDS: *proteomics, digital library, bibliometric, text-mining, information retrieval, artificial intelligence, novelty detection, citation.*

1. Introduction

De nombreuses techniques aident les scientifiques à rechercher et obtenir des informations appropriées dans une collection documentaire. La plupart de ces systèmes, qui utilisent un modèle d'interaction utilisateur appelé recherche d'information *ad hoc,* fournissent à l'utilisateur des masses de documents de divers degrés de similitude entre la requête fournie en entrée et les documents retournés, le plus souvent classés selon un ordre croissant de pertinence. Ce modèle, popularisé par des outils de recherche sur le Web, tels que Google ou Alta Vista, s'est clairement établi comme la tâche dominante en recherche d'information. Cependant, les critiques de ces systèmes sont nombreuses et, dans cet article, nous nous attachons à étendre ce modèle en nous proposant de définir une architecture de recherche d'information capable de détecter la présence d'un nouveau document pertinent pour une requête donnée. L'application que nous visons est la mise à jour d'information dans une base de données de protéomique (Bairoch 1997). De telles bases de connaissances, dont UniProtKB/Swiss-Prot est l'une des plus importantes (Wu et al., 2006), utilisent massivement l'information bibliographique telle que disponible dans la bibliothèque digitale MEDLINE (Bourne 2006, Grivell 2002), pour mettre à jour son contenu.

Le problème de la recherche d'informations nouvelles (ou détection de nouveautés) a été relativement peu étudié en recherche d'information. On trouve par exemple une tentative dans ce sens dans le cadre de la « *TREC Novelty track* » (Soboroff et Harman 2003). La tâche proposée était formellement définie comme une tâche de recherche de passage, ou plus précisément, comme une tâche de recherche de phrases qu'il convenait d'ordonner les unes par rapport aux autres. Différentes méthodes, le plus souvent basées sur des techniques mélangeant une pondération statistique (Collins-Thompson et al., 2002) et une exploitation de patrons spécifiques (Li et Croft TREC 2002) ciblant un ensemble d'entités nommées comme proposé par les systèmes de question-réponse (Soboroff et Harman 2003).

Notre approche est originale dans la mesure où nous essayons d'utiliser de l'information bibliométrique combinée à de la recherche d'information classique pour détecter de l'information à la fois pertinente et nouvelle. Nous nous proposons d'utiliser les références bibliographiques utilisées pour décrire une protéine, afin de découvrir des articles utiles pour la mise à jour de l'information concernant cette protéine. Nous faisons l'hypothèse qu'un article citant un article déjà utilisé pour décrire la protéine est un article potentiellement intéressant.

1.1. *Fouille de données et bibliométrie*

La fouille de données textuelles est définie comme une technique qui permet la découverte d'informations cachées dans des contenus textuels. Grâce à ces

techniques, on peut découvrir une suite de liens logiques cruciaux dans une grande collection de documents (par exemple, la littérature biomédicale) qu'il serait difficile de traiter par un humain avec des méthodes traditionnelles. La plupart des tâches de fouilles de données textuelles débutent par une tâche de recherche d'information (RI), ou recherche documentaires (RD).

Provenant de la bibliométrie, l'analyse de références bibliométriques (White 2003, Liu 1993) a été employée pour visualiser un domaine d'étude par l'intermédiaire d'une tranche représentative de sa littérature. Les techniques utilisant des réseaux de co-citation permettent de grouper des documents par paradigme scientifique (Noyons et al. 1999). Braam et al. (1991) ont étudié les co-citations comme un outil pour catégoriser automatiquement les spécialités des domaines scientifiques. Ils ont constaté que la combinaison de l'analyse de mot-clé et de l'analyse de co-citations était utile pour indiquer la teneur cognitive des publications.

Peters et al. (1995) sont allés plus loin, ils ont exploré les relations de citations et la ressemblance cognitive dans les articles scientifiques. Les similitudes des profils des mots des publications qui ont été bibliographiquement couplées à un article fortement cité ont été comparées aux publications qui n'ont pas été bibliographiquement couplées à cet article spécifique. Une relation statistiquement significative a été établie entre le contenu des articles et leurs citations partagées (Wu et Crestani, 2003 ; Soborrof et al. 2001). Cette propriété a notamment été exploitée pour générer automatiquement des jugements de pertinences dans une tâche de recherche d'articles similaires dans MEDLINE (Tbahriti et al. 2005).

1.2. *Base de connaissance Swiss-Prot*

Il s'agit d'une base de données de séquences de protéines avec un niveau élevé d'annotations, telle que la description de la fonction d'une protéine, de sa structure, de ses variants et des maladies pouvant être causé par ces variants. La figure 1 montre un exemple d'une entrée UniProtKB/Swiss-Prot avec quelques unes des lignes (champs) qui la composent. Ces annotations sont générées par une équipe d'une centaine de biologistes qui recherchent et compilent les connaissances sur une protéine donnée en fouillant la littérature publiée sur cette dernière. Les articles qui, par ce biais, ont été utilisés pour produire l'annotation sont référencés dans l'entrée de la protéine.

On peut schématiquement décrire la tâche d'annotation d'UniProtKB/Swiss-Prot comme une tâche de catégorisation dans un vocabulaire contrôlé comme la « Gene Ontology » (GO). Différentes études, par exemple dans le cadre de la campagne d'évaluation BioCreative (Blaschke et al., 2005), ont été mises en place afin d'aider les annotateurs dans leurs tâches. En plus de ces tâches de classification pure, dans la Gene Ontology ou dans d'autres systèmes terminologiques (Ruch 2006), l'annotation en protéomique fonctionnelle requiert également des tâches

d'extraction de passage (Ehrler et al., 2005). D'autres tâches d'annotation de protéines s'apparentent à des tâches de génération de résumés automatiques (Ruch 2005) ou de filtrage (Dobrokhotov et al. 2003). Enfin, la plupart de ces tâches présupposent une recherche d'information efficace dans la bibliothèque digitale MEDLINE (Abdou et al. 2005), comme investiguée par la compétition « TREC Genomics » (Hersh 2003).

```
ID   CFTR_HUMAN              Reviewed;       1480 AA.
AC   P13569;
DT   12-DEC-2006, entry version 103.
DE   Cystic fibrosis transmembrane conductance regulator (CFTR) (cAMP-
DE   dependent chloride channel) (ATP-binding cassette transporter sub-
GN   Name=CFTR; Synonyms=ABCC7;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OX   NCBI_TaxID=9606;
…
RX   MEDLINE=89368940; PubMed=2475911;
RL   Science 245:1066-1073(1989).
CC   -!- FUNCTION: Involved in the transport of chloride ions. May regulate
CC       bicarbonate secretion and salvage in epithelial cells by
CC       regulating the SLC4A7 transporter.
…
CC   -!- DISEASE: Defects in CFTR are the cause of congenital bilateral
CC       absence of the vas deferens (CBAVD) [MIM:277180]. CBAVD is an
CC       important cause of sterility in men and could represent an
CC       incomplete form of cystic fibrosis, as the majority of men
CC       suffering from cystic fibrosis lack the vas deferens.
...
DR   EMBL; M28668; AAA35680.1; -; mRNA.
DR   EMBL; M55131; AAC13657.1; -; Genomic_DNA.
KW   3D-structure; ATP-binding; Chloride; Chloride channel;
FT   VARIANT      13     13       S -> F (in CF).
FT                                /FTId=VAR_000101.
SQ   SEQUENCE   1480 AA;  168174 MW;  8D082BF6B628D065 CRC64;
     MQRSPLEKAS VVSKLFFSWT RPILRKGYRQ RLELSDIYQI PSVDSADNLS
     EKLEREWDRE…
```

Figure 1. *Exemple d'une entrée Swiss-Prot partielle, l'entrée contient, entre autre, un (ID) et un numéro d'accession unique (AC), sa date de création (DT), le nom complet de la protéine, avec acronymes et synonymes (DE), le nom du gène correspondant (GN), ainsi que l'organisme dans lequel la séquence a été identifiée (OS,OG), une liste de références (R*), du texte libre décrivant les caractéristiques de la protéine (CC), ainsi que les propriétés de la séquence (FT), cette information est complétées par des cross-références d'autres bases de données (DR).*

Dans la suite de l'article, la section 2 détaille le développement et les méthodes utilisés, la section 3 présente les résultats de l'expérience et leur évaluation et pour finir la section 4 termine avec une conclusion, les limitations et les travaux futurs.

2. Méthodes

Comme schématisé dans la figure 2, nous utilisons deux différentes manières pour mettre en œuvre notre expérience. Nous définissons deux méthodes, la première est guidée par les références (GR), la seconde basée sur la recherche documentaire (GD). Dans les deux cas, le point de départ est la base de données Swiss-Prot et parmi l'ensemble des références de base (ERB) répertoriées pour chaque protéine, nous utilisons pour tester les méthodes, un sous-ensemble comme cibles (ERC), et un sous-ensemble comme liste de références à découvrir (ERD). Le système sera considéré optimal s'il est capable de retourner un ensemble de références possibles (ERP) identique à l'ensemble de références à découvrir (ERD).

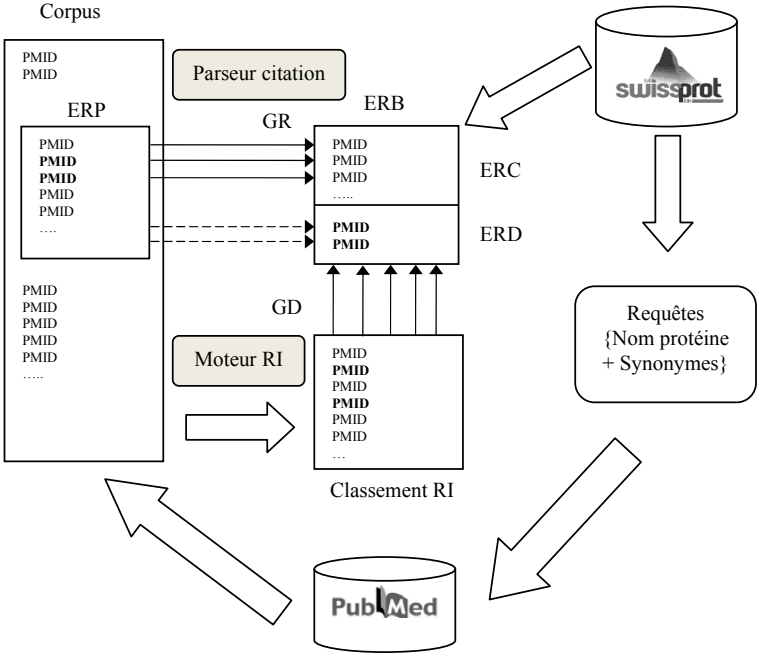

Figure 2. *Organigramme pour la chaîne des procédures expérimentales.*

2.1. *Métriques*

Pour implémenter la méthode basée sur les réseaux de citations (GR), la première étape consiste à déterminer l'ensemble des références candidates possibles (ERP), grâce à l'analyse de leurs citations. Par la suite, la deuxième étape, consiste à générer, à partir d'ERP, des classements basés sur le nombre de références citant des éléments de l'ensemble des références bibliographiques de l'entrée Swiss-Prot (ERB). En d'autres termes, plus un article contiendra de références bibliographiques apparaissant dans ERB, plus on considérera que l'article est susceptible d'être utile pour mettre à jour l'information de la protéine considérée.

Le moteur de recherche génère aussi un classement des articles en se basant sur le calcul d'une distance entre les caractéristiques présentent dans la requêtes et celles présente dans les documents retourné ; le calcul de cette distance est fourni plus dans l'équation [1]. Le réglage du schéma de pondération du moteur de recherche utilisé pour générer nos résultats en se basant sur la recherche documentaire (GD), s'appuie sur le résultat d'expériences antérieures (Aronson, TREC 2005). Dans ces expériences, on maximise en général la précision moyenne (*Mean Average precision* ou MAP) et nous faisons l'hypothèse qu'un tel réglage est bon pour la tâche qui nous intéresse présentement. MAP est, en effet, la métrique la plus communément utilisée en RI, bien qu'elle tende parfois à cacher des différences mineures dans le classement des documents (Mittendorf et Schäuble, 1994).

Dans l'équation [1], le poids de la requête et du document est donné; le premier triplet (*dtu*) s'applique au document, pendant que le deuxième triplet (*dtn*) est appliqué à la requête ; t représente le nombre de termes indexés, df_j, le nombre de documents où le terme t_j apparaît ; pivot et slope sont des constantes déterminées empiriquement par des expériences antérieures (pivot = 0.14, slope = 146), cf. Aronson et al. (2005) pour une présentation plus détaillée.

$$\text{dtu} : w_{ij} = \frac{(Ln(Ln(tf_{ij})+1)+1).idf_j}{(1-slope).pivot + slope.nt_i}$$

$$\text{dtn} : w_{ij} = idf_j.(Ln(Ln(tf_{if})+1)+1) \qquad [1]$$

Afin de comparer les deux méthodes, celle basée sur la recherche documentaire fournie par le moteur de recherche et celle fournie pas le nombre de citations, on calcule une corrélation selon la méthode Pearson, sans tolérance des valeurs omises. Ensuite, on essaie d'analyser qualitativement cette corrélation entre les deux classements, afin de vérifier s'il s'agit d'une simple redondance ou plutôt d'une complémentarité entre les deux méthodes. Enfin, nous essaierons de fusionner les deux méthodes par combinaison linéaire, selon l'approche combSUM (Shawn et Fox, 1994).

2.2. *Acquisition de données et indexation des citations*

Un ensemble de 50 entrées UniProtKB/Swiss-Prot ont été choisies aléatoirement dans la base de données. On s'assure que chacune de ces protéines dispose de plus de deux références bibliographiques dans MEDLINE afin de pouvoir séparer l'ensemble de références de chaque protéines (ERB) en deux : les références dont on fait l'hypothèse qu'elles sont connues (ERC), et celles à découvrir par le système (ERP). Le contenu des références, c'est-à-dire le résumé, le titre et les mots clefs d'indexation, sont récoltés via MEDLINE en utilisant l'interface PubMed.

2.2.1. *Requêtes et collection de documents*

Afin de collecter de nouveaux documents correspondants aux entrées UniProtKB/Swiss-Prot décrites précédemment, nous construisons des requêtes booléennes disjonctives simples basées sur le nom de la protéine et de ces synonymes. Dans certains cas, le volume de la collection récolté est très grand, et nous devons restreindre la liste des synonymes de la protéine en excluant des synonymes trop génériques et en rajoutant à la requête le nom de l'organisme dans lequel la protéine a été séquencée. Nous récupérons ainsi un total de 65000 références. On remarque que le nombre de documents varie significativement d'une requête à l'autre : de 29 à 14000, et en moyenne 2700 références, avec un écart-type important de 2900.

Pour nos besoins de catégorisation, nous ne conservons du contenu des documents que le titre, le résumé, les termes MeSH (Medical Subject Headings) et les noms chimiques. Nous fournissons ensuite ces documents au système de la recherche d'informations pour indexation. On préfère utiliser ces champs de MEDLINE plutôt que le texte intégral des articles afin de réduire la taille des corpus à manipuler. De plus, de nombreux travaux tendent à montrer que l'utilisation du contenu intégral des articles n'est pas préférable à l'utilisation des résumés pour des tâches de recherche d'information dans MEDLINE. En particulier, Schuemie et al., (2004) montrent que les résumés contiennent une densité plus élevée d'informations que les textes intégraux. Dans le même ordre d'idée, Gay et al. (2005) montre que l'utilisation de l'ensemble de l'article n'apporte qu'un gain modeste et non significatif dans le cadre d'une tâche de catégorisation automatique.

La récupération des textes intégraux est cependant nécessaire pour la méthode guidée par les références (GR), puisque la bibliographie des articles devra être intégralement analysée. Toutefois, nous ne pouvons récupérer le texte intégral des articles au format HTML que pour 24000 de ces références. Les références bibliographiques de ces articles sont automatiquement extraites et rattachées aux identificateurs PubMed (PMID) correspondants. Evalué sur un échantillon de 80 articles, notre analyseur arrive à extraire les références des articles avec une précision de 95%, malgré la difficulté de la tâche due à l'hétérogénéité entre différents journaux. Pour les 5% restants, l'analyseur n'arrive pas à rattacher la

citation à un PMID soit parce que le format de la citation n'est pas conforme (maque d'information, par exemple le volume), soit la citation elle-même n'existe pas dans MEDLINE.

2.2.2. *Indexation des textes*

Pour l'indexation, nous employons le système EasyIR[1], qui implémente un modèle vectoriel standard pour la recherche d'information (Salton et al. 1983). Il propose la plupart des schémas de pondération *tf.idf* classiques et quelques schémas plus avancés (normalisation à pivot, Okapi). Comme pour le moteur SMART dont il s'inspire, les schémas de pondération sont représentés sous forme de paires de triplets : document-requête. Le triplet représente respectivement, la fréquence du terme dans le document (*term frequency* ou TF), sa fréquence dans la collection (*Inverse document frequency* ou IDF), et un facteur de normalisation. Toutes nos expériences sont entreprises avec une normalisation pivotée (Singhal et al., 1996a), qui semple récemment montrer une certaine efficacité pour des tâches de RI dans MEDLINE (Aronson et al., TREC 2005).

3. Résultats et discussion

Dans cette section, nous comparons les résultats obtenus par chacune des deux méthodes et discutons qualitativement des résultats respectifs de chacune des méthodes sur quelques exemples sélectionnés. Nous faisons l'hypothèse que l'ensemble de document retourné par le moteur de recherche (guidage par la recherche documentaire) fournit une bonne base de départ pour l'évaluation et la comparaison de l'approche guidée par les références bibliographiques. Dans un premier temps, nous détaillons quelques exemples avant de nous intéresser aux résultats d'ensemble.

3.1. *Exemple de comparaison simple*

Le tableau 1 donne trois exemples de protéines (l'identifiant Swiss-Prot est utilisé) distribués selon le nombre de références répertoriées (ERB), cibles (ERC), à découvrir (ERD), découverte par la méthode guidée par les références (GR), découverte par la méthode guidée par la recherche documentaire (GD). Pour chaque protéine utilisée dans l'étude, on a déterminé le nombre de références à découvrir : en général, on fixe le nombre de référence à découvrir à un quart du nombre total de références (arrondi à l'inférieur), mais quand le nombre de référence total est inférieur à quatre, le nombre de référence à découvrir est limité à une unique

[1] http://www.natlang.hcuge.ch/Resources/softs.htm

référence. Pour la méthode guidée par la recherche documentaire, le nombre de références retournées est limité à 1000.

	ERB	***ERC***	***ERD***	***GR***	***GD***
P30429	3	2	1	1	1
Q9UJY5	21	16	5	5	5
O61967	2	1	1	-	1

Tableau 1. *Exemple de la validation des méthodes.*

Dans le tableau 1, la protéine UniProtKB/Swiss-Prot P30429 cite trois références dans MEDLINE (ERB). Parmi ces références, le sous-ensemble ERD des références à découvrir contient une unique référence. La méthode guidée par les références (GR) est capable de la trouver en utilisant les deux références restantes (ERC). De même la méthode basée sur la recherche documentaire (GD), en utilisant simplement le nom de la protéine et ses synonymes semble capable d'un résultat similaire. Pour la protéine dont l'identifiant est Q9UJY5, avec 21 références dans ERB, les deux méthodes sont capables de produire des résultats comparables : les 5 références à découvrir (ERD) le sont par chacune des deux méthodes. En revanche, pour la dernière protéine (O61967), seule la méthode GD est capable de découvrir l'unique référence attendue.

Bien que dans les exemples du tableau 1, la méthode guidée par la recherche documentaire semble donner de meilleurs résultats, il faut remarquer que les méthodes ne sont pas directement comparables à l'aide des scores fournis dans les colonnes GR et GD. En effet, pour une comparaison plus fine, il est nécessaire de considérer à quel rang les références à découvrir sont proposées par chacune des méthodes.

3.2. *Exemple de comparaison des classements*

3.2.1. *Classements équivalents*

Le tableau 2 montre un exemple où les classements des deux méthodes sont presque équivalents. Par exemple, la première référence de l'entrée UniProtKB/Swiss-Prot, appartenant à ERB (PMID:15383288) est détectée par les deux méthodes : la méthode guidée par les références obtient l'article en tête de liste, avec trois références citée dans ERB (colonne GR), alors que la méthode guidée par la recherche documentaire le retourne en septième position (colonne GD).

Collection	ERB	GR	GD
15383288 (2004)	**X**	**3**	**7**
15238520 (2004)		2	14
15048082 (2004)		1	23
14993223 (2004)		1	1
14647239 (2004)		1	20
12963020 (2003)		1	5
12563030 (2003)		1	3304
12556527 (2003)		1	19
15919588 (2005)		0	2263
15914584 (2005)		0	1039
15907692 (2005)		0	1839
15906142 (2005)		0	2058

Tableau 2. *Exemple d'un classement où les deux méthodes sont relativement équivalentes. La première colonne donne l'identifiant de l'article (PMID) dans MEDLINE. La seconde colonne indique si l'article considéré est une référence citée dans SwissProt. La colonne trois, donne le nombre de références apparaissant dans l'enregistrement de la protéine et cité dans l'article. La colonne quatre, donne le rang de l'article tel que retourné par le moteur de recherche, c'est-à-dire la méthode basée sur la recherche documentaire.*

3.2.2. *Classements hétérogènes*

Dans les tableaux suivants (tableau 3 et 4), on trouve un exemple où les deux méthodes ont des classements très différents. Par exemple, le tableau 3 montre les références (14578922, 10878806) appartenant à l'ensemble ERB qui sont trouvées par la méthode GD, mais pas par la méthode GR. Une autre référence (11493666) qui est ignorée par GR mais classée candidat numéro un par GD.

Le tableau 4, montre l'inverse, c'est-à-dire des articles (par exemple 12145329, 12650700) qui citent beaucoup de références dans ERB, donc très pertinents dans la méthode GR, et qui se trouvent parmi les derniers dans le classement de la méthode GD. Une autre référence aussi (12137800) qui est candidate numéro un dans GR mais avec un rang très élevé dans GD, donc ignorée.

Collection	ERB	GR	GD
15806148 (2005)		2	7
15063180 (2004)		2	4
14729475 (2004)		1	9
11711544 (2002)		1	8
9442884 (1997)		0	52
9305626 (1997)		0	17
9287107 (1997)		0	36
9182656 (1997)		0	30
8139009 (1994)		0	32
7721789 (1995)		0	22
15767459 (2005)		0	20
15569358 (2004)		0	48
15175862 (2004)		0	58
15010317 (2004)		0	37
14578922 (2003)	**X**	**0**	**6**
12961049 (2003)		0	10
12867631 (2003)		0	59
12706335 (2003)		0	38
11821050 (2002)		0	53
11792814 (2001)		0	65
11715019 (2001)		0	3
11607032 (2001)		0	41
11493666 (2001)		**0**	**1**
11257190 (2001)		0	29
11161560 (2001)		0	11
10878806 (2000)	**X**	**0**	**2**

Collection	ERB	GR	GD
12137800 (2002)		**12**	**2776**
12754700 (2003)		9	31
12145329 (2002)		**9**	**2642**
14749509 (2004)		**7**	**1241**
12650700 (2002)		**7**	**2705**
10766198 (2000)		**7**	**2257**
15469870 (2004)		**6**	**1224**
15895983 (2005)		5	782
12943736 (2003)		5	2135
12920235 (2003)		5	2333
11932321 (2002)		5	1664
11589680 (2001)	**X**	**5**	**1100**
10601116 (1999)		5	16
9920778 (1999)		4	513
15770078 (2005)		4	35
15488991 (2004)		4	2969
15017388 (2004)		4	1883
14724649 (2004)		4	1028
12954798 (2003)		4	1858
12623935 (2003)		4	4
12431798 (2002)		4	2623
12088883 (2002)		4	1194
12050230 (2002)	**X**	**4**	**111**
11701741 (2001)	**X**	**4**	**1396**
10843199 (2000)		4	1325
10481166 (1999)		4	2887
10096574 (1999)		4	1158

Tableaux 3-4. *Exemple de classements hétérogènes*

3.3. *Corrélation*

On mesure la corrélation dans l'intervalle [0, 1], plus la valeur se rapproche de un, plus on a une forte corrélation et inversement.

Sur la figure 3, on observe assez clairement la corrélation qui existe entre les deux méthodes pour l'entrée P27129. Les articles ayant cités plus de références Swiss-Prot (ERB) ont tendance à être placés en tête du classement GR et en tête de classement par le moteur de recherche d'information (classement RI). Ils sont donc considérés comme les plus intéressants par les deux méthodes.

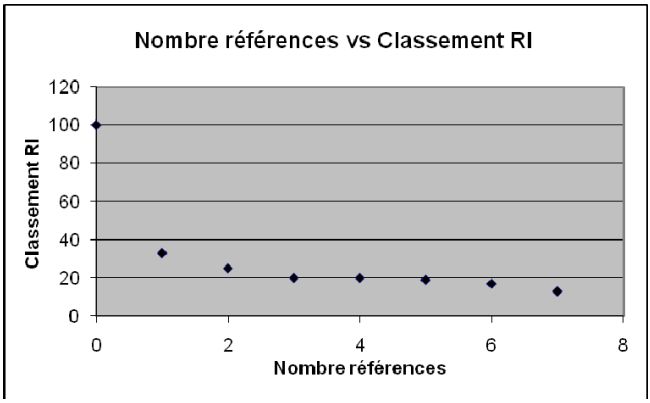

Figure 3. *Exemple de corrélation des classements (entrée P27169) entre le rang retourné par le moteur de recherche (méthode guidée par la recherche d'information) et le rang obtenu par comptage du nombre de références citées dans l'enregistrement de la protéine.*

Le tableau 5, donne la corrélation des classements de l'ensemble des cinquante protéines UniProtKB/Swiss-Prot utilisées dans nos expériences. On donne également quelques exemples des corrélations observées pour six protéines.

Protéines	Corrélations
O34483	0,8326432
O61967	0,9111897
P00750	0,9990153
P04150	0,6589558
P30429	0,898355
P35568	0,7892328
Moyenne (50 protéines)	**0,74**
Ecart-Type	0,35

Tableau 5. Corrélation des classements entre les deux méthodes.

Pour comparer la précision des deux méthodes et dans la mesure où les deux méthodes ne retournent pas le même nombre de documents, on peut mesurer la précision au rang 1, c'est-à-dire la précision de la première référence retournée ou P0, par chacune des méthodes (cf. Tableau 6). On observe que la méthode guidée par la recherche d'information propose à peu près une fois sur deux (0.46%) un article attendu, tandis que la méthode guidée par les références se comporte moins bien, 0.31%, soit une fois sur trois seulement. Ce qui semble suggérer que la capacité d'ordonnancement du modèle basé sur la recherche documentaire est supérieure à celle du système guidé par les références. Enfin, dans le Tableau 6, la fusion des deux méthodes, par simple combinaison linéaire, montre un gain significatif de 6.5% ($p<0.01$) par rapport à la méthode basée sur la recherche documentaire, que l'on prend comme référence de base. Ce dernier résultat permet de confirmer la complémentarité et l'hétérogénéité relative de chacune des méthodes pour une tâche de détection de nouveaux documents pertinents.

	P0 (%)
Guidage par les références	0,31
Guidage par la recherche documentaire (référence)	0,46
Combinaison	0.49 (+6.5%)

Tableau 6. *Précision (P0) comparée de chaque méthode et de leur combinaison.*

4. Conclusion et travaux futurs

Nous avons développés une méthode, pour l'identification d'articles utiles à la mise à jour de l'information dans une base de données de biologie moléculaire, qui utilise un analyseur de bibliographie afin de détecter des articles nouveaux citant des articles connus. Cette méthode est comparée à une méthode utilisant un moteur de recherche. On remarque que les deux méthodes sont bien corrélées, mais que la méthode basée sur la recherche documentaire semble fournir une meilleure performance au regard de la précision dans les premiers documents retournés. La combinaison de chacune des méthodes apporte également un gain significatif de précision par rapport à chacune des méthodes séparément. Cette complémentarité permet d'envisager différentes stratégies de fusion afin de produire des systèmes de détection de nouveautés plus performants. Il faut aussi remarquer que les deux méthodes proposent des documents non référencés dans la base de données UniProtKB/Swiss-Prot. Il serait donc nécessaire de vérifier la qualité de ces références du point de vue de l'expert, en particulier dans une tâche de détection de nouvelles connaissances pour l'annotation d'UniProtKB/Swiss-Prot. Une validation manuelle des deux méthodes, effectuée par deux annotateurs UniProtKB/Swiss-Prot, est actuellement en cours.

Enfin, la topologie des réseaux de citations est formellement similaire à ceux de graphes imbriqués, il serait donc intéressant d'investiguer comment des algorithmes basés sur ces approches (Brin et Page, 1998), pourraient complémenter la recherche d'information classique dans des collections exhibant des propriétés similaires, comme c'est le cas des bases de données bibliographiques.

Bibliographie

S. Abdou, J. Savoy, P. Ruch: Dépister efficacement de l'information dans une banque documentaire : L'exemple de MEDLINE. *INFORSID* 2006: p. 129-143.

A.R. Aronson, D. Demner-Fushman, S.M. Humphrey, J. Lin, H. Liu, P. Ruch, M.E. Ruiz, L.H. Smith, L.K. Tanabe, W.J. Wilbur, « Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. », *TREC 2005*, Gaithersburg, MD, USA.

A. Bairoch, *Proteome Databases in Proteome Research*, M. R. Wilkins, K. L. Williams, R. D. Appel, D. F. Hochstrasser (Eds.), Berlin, Springer, 1997.

C. Blaschke, E. A. Leon, M. Krallinger, A. Valencia, « Evaluation of BioCreAtIvE assessment of task 2 », *BMC Bioinformatics,* vol. 6, Suppl. 1, 2005, p. S16.

P. Bourne, « Will a biological database be different from a biological journal? », *PLoS Comput Biol*, vol. 1 n° 3, Aug 2005, p. 179-81.

R.R. Braam, H.F. Moed, A.F.J. Van Raan, « Mapping of science by combined co-citation and word analysis. Part I; Structural aspects », *J. Am. Soc. Inform. Sci.,* vol. 42, n° 4, 1991, p. 233-251.

S. Brin et L. Page, « The anatomy of a large-scale hyper textual Web search engine », *Université de Stanford*, 1998.

K. Collins-Thompson, P. Ogilvie, Y Zhang, J. Callan, « Information Filtering, Novelty Detection, and Named-Page Finding », *TREC 2002.*

P. Dobrokhotov, C. Goutte, A-L. Veuthey, E. Gaussier. A Probabilistic information retrieval approach to medical annotation in SWISS-PROT. Medical informatics Europe. The New Navigators, R. Baud, M. Fieschi, P. Le Beux, and P. Ruch (Eds), 2003, IOS Press.

F. Ehrler, A. Geissbühler, A. Jimeno, P. Ruch, « Data-poor categorization and passage retrieval for Gene Ontology Annotation in Swiss-Prot », *BMC Bioinformatics,* vol. 6, Suppl. 1, 2005, p. S23.

C.W. Gay, A.R. Aronson, M. Kayaalp, « Semiautomatic indexing for online biomedical journals. », *AMIA 2005.*

L. Grivell, « Mining the bibliome: searching for a needle in a haystack? New computing tools are needed to effectively scan the growing amount of scientific literature for useful information. », *EMBO Rep*., vol. 3, n° 3, 2002, p. 200-3.

W.R. Hersh, R.T. Bhupatiraju « TREC GENOMICS Track Overview », *TREC 2003.*

X. Li, W. Bruce Croft, « Novelty Detection Based on Sentence Level Patterns », *TREC 2002*.

M. Liu, « The complexities of citation practice : a review of citation studies », *Journal of Documentation*, vol. 49, n°4, 1993, p. 370-408.

E. Mittendorf, P. Schäuble, « Document and Passage Retrieval Based on Hidden Markov Models. », *SIGIR 1994*, p. 318-327.

E.C.M. Noyons, H.F. Moed, M. Luwel, « A bibliometric study combining mapping and citation analysis for evaluative bibliometric purposes », *J. Am. Soc. Inform. Sci.*, vol. 50, n° 2, 1999, p. 115-131.

H.P.. Peters, R.R. Braam, A.F.J. Van Raan, « Cognitive resemblance and citation relations in chemical engineering publications », *J. Am. Soc. Inform. Sci.,* vol. 46, n° 1, 1995, p. 9-21.

D. Rebholz-Schuhmann, H. Kirsch, F. Couto, « Facts from text-is text mining ready to deliver? », *PLoS Biol.*, vol. 3, n° 2, 2005 Feb, p. e65.

P. Ruch, « Automatic assignment of biomedical categories: toward a generic approach », *Bioinformatics,* vol. 22, 2006, p. 658-664.

P. Ruch, L. Perret, J. Savoy, « Features Combination for Extracting Gene Functions from MEDLINE », *ECIR 2005*, p. 112-126, Springer LNCS.

G. Salton, M.J. McGill, « Introduction to Modern Information Retrieval. », *McGraw-Hill*, 1983.

M.J. Schuemie, M. Weeber, B.J.A. Schijvenaars, E.M. van Mulligen, C.C. van der Eijk, R. Jeliert, B. Mons, J.A. Kors, « Distribution of information in biomedical abstracts and full text publications », *Bioinformatics*.

J. Shaw and E. Fox, Combination of Multiple Searches, *TREC*, 1994.

A. Singhal, C. Buckley, M. Mitra, « Pivoted Document Length Normalization. », *ACM SIGIR*, 1996, p. 21-29.

I. Soborrof, C. Nicholas, P. Cahan, «Ranking retrieval systems without relevance judgments», *SIGIR*, 2001, p. 66-73.

I. Soboroff, D. Harman, « Overview of the TREC 2003 Novelty Track », *TREC 2003.*

I. Tbahriti, C. Chichester, F. Lisacek, P. Ruch, « Using argumentation to retrieve articles with similar citations: an inquiry into improving related articles search in the MEDLINE digital library », *Int J Med Inform.,* vol. 75, n° 6, Jun 2006, p. 488-95.

H. White, « Pathfinder networks and author co-citation analysis: a remapping of paradigmatic information scientists », *J. Am. Soc. Inf. Sci. Technol.*, vol. 54, n° 5, 2003, p. 423-434.

C. Wu, R. Apweiler, A. Bairoch, D. Natale, W. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, B. Suzek, « The Universal Protein Resource (UniProt): an expanding universe of protein information », *Nucleic Acids Res*, vol. 34, 2006, p. D187-D191.

S. Wu, F. Crestani, « Methods for Ranking Information Retrieval Systems without Relevance Judgments », *SAC*, 2003, p. 811-816.