Extraction d'information dans les bases de données textuelles en génomique au moyen de transducteurs à nombre fini d'états

Thierry Poibeau

Thales Recherche et Technologie
Domaine de Corbeville
91404 Orsay

Laboratoire d'Informatique de Paris-Nord
avenue J.-B. Clément
93340 Villetaneuse

Thierry.Poibeau@lcr.thomson-csf.com

Résumé – Abstract

Cet article décrit un système d'extraction d'information sur les interactions entre gènes à partir de grandes bases de données textuelles. Le système est fondé sur une analyse au moyen de transducteurs à nombre fini d'états. L'article montre comment une partie des ressources (verbes d'interaction) peut être acquise de manière semi-automatique. Une évaluation détaillée du système est fournie.

This papers describes a system extracting information about interactions between genes or proteins, from large textual databases. The system is based on a set of linguistic finite-state transducers. The paper shows how a part of the resources (namely the set of verbs expressing the notion of interaction) can be acquired semi-automatically from the corpus. A detailed evaluation is provided.

Mots clés : extraction d'information, génomique, transducteurs linguistiques

1 Introduction

Depuis une dizaine d'années sont apparues de nombreuses bases de données concernant la biologie. Ces bases de données sont de plusieurs types : banques de séquences nucléiques ou protéiques, banques de structures moléculaires, bases de données génomiques (d'après (Pillet, 2000). Ce dernier type est actuellement le plus étudié, du fait des enjeux en cours autour du déchiffrage du génome et de ses conséquences pour la recherche médicale. Les bases génomiques peuvent elles-mêmes être divisées en deux grandes familles, selon qu'elles sont généralistes (par exemple Medline) ou spécialisées (Flybase, consacrée à la drosophile).

Ces bases de données ont été constituées progressivement, au fur et à mesure que le déchiffrement du génome progressait. Les bases sont donc naturellement structurées par gènes, les informations concernant les différents gènes étant régulièrement mises à jour en

fonction des publications, brevets et autres informations rendues publiques. Le véritable enjeu pour la recherche en génomique est toutefois ailleurs : maintenant que la phase de déchiffrage se termine, une autre commence qui vise à comprendre le fonctionnement correct des cellules via l'étude des interactions génétiques. Mais les chercheurs sont alors confrontés à des bases structurées par gènes, alors qu'ils cherchent des interactions entre gènes. Cette information figure bien dans les bases visées mais elle est éparpillée sous forme textuelle. Le problème est énorme quand on sait par exemple que Medline est actuellement composé au total de plus de seize millions de notices.

Cet article présente des méthodes et des outils pour l'extraction de tables d'interaction entre gènes à partir de bases de données textuelles de génomique. Après avoir dressé un rapide état de l'art du domaine, nous présentons différentes réalisations effectuées sur Flybase. Nous montrons comment les notices sur les gènes doivent être « préparées » afin de cibler la recherche sur les éléments pertinents. Nous présentons ensuite des techniques pour le repérage des noms de gènes, puis des interactions entre gènes. Nous verrons à cette occasion que les techniques employées se rapportent à des techniques utilisées en terminologie et en extraction d'information. Nous présentons enfin une évaluation et les perspectives de cette étude.

2 Travaux antérieurs

La recherche d'information au sein de bases de données textuelles biologiques a donné lieu à de nombreuses études ces dernières années. Citons par exemple (Goujon, 1999) qui propose un système fondé sur l'analyse contextuelle pour la recherche d'information au sein de notices sur les plantes transgéniques. Un jeu de marqueurs pertinents est défini pour pouvoir faciliter la lecture des notices en mettant en surbrillance les passages *a priori* les plus importants.

Plusieurs travaux, anglo-saxons pour la plupart, ont porté sur l'analyse de bases de données en génomique pour en extraire des informations sur les interactions entre gènes. A côté de nombreux travaux reposant sur une approche stochastique (par exemple, (Stapley *et al.* 2000) établissent des cartes à partir de l'analyse de cooccurrences de noms de gènes) commencent à apparaître des travaux fondés sur une analyse linguistique. La recherche d'interaction pouvant être assimilée au remplissage d'un formulaire d'extraction, deux études récentes ont proposé d'adapter des systèmes d'extraction d'information déjà existants. Il s'agit de l'étude de (Thomas *et al.*, 2000) qui adapte le système d'extraction Highlight développé à Cambridge, et de (Humphreys *et al.*, 2000) qui adapte le système LaSIE développé à l'Université de Sheffield. Les résultats obtenus sont intéressants mais l'effort d'adaptation de tels systèmes reste à évaluer. Une approche plus locale et plus ciblée est peut-être mieux indiquée pour la tâche que celle proposée par (Humphreys *et al.*, 2000), qui implique la mise en place préalable d'une ontologie et d'un « modèle du discours » (*discourse model*, pour reprendre les termes employés par les auteurs).

Pillet (2000) a proposé dans sa thèse une « méthodologie d'extraction automatique d'information » à partir de bases en génomique « en vue d'alimenter un système d'information ». L'auteur a développé, à partir de Flybase, une stratégie de filtrage des phrases potentiellement pertinentes, de marquage des noms de gènes puis vérifié manuellement qu'il était effectivement question d'interaction dans ces phrases. Le but est

ensuite d'appliquer des méthodes statistiques (analyse factorielle des données et index de vraisemblance d'interaction, IVI) pour prédire qu'une phrase mentionne ou non une interaction, à partir de l'apparition d'un vocabulaire spécifique mis en évidence sur le corpus de référence validé manuellement. La méthode présente l'inconvénient de masquer les interactions peu décrites mais souvent précieuses ; elle ne permet pas non plus de dire la nature du lien entre les deux gènes, ni le sens de la relation.

Nous proposons à notre tour une méthode qui s'inspire des travaux de Pillet pour le filtrage des données mais qui, pour le reste, fait intervenir des traitements linguistiques, à l'instar de (Thomas *et al.*, 2000) et de (Humphreys *et al.*, 2000). Le corpus est d'abord filtré et structuré au moyen de programmes Perl puis analysé par un ensemble de transducteurs linguistiques, élaborés grâce à la boîte à outils Intex (Silberztein, 1993).

3 Description du corpus et préparation des données

Le corpus a été constitué suite à l'interrogation de Flybase. Un ensemble de cent notices (chacune se rapportant à un gène donné) a servi à la mise au point du système et cent autres ont été réservées pour l'évaluation. Une notice est un ensemble d'information structuré en plusieurs champs. Par manque de place nous ne pouvons reproduire un extrait du corpus mais l'étude de Pillet (2000) donne une description détaillée des notices Flybase. Dans ce type de base textuelle, l'information pertinente est noyée dans un océan de texte (une notice peut comprendre plusieurs dizaines de pages imprimées).

Le corpus décrit en effet beaucoup d'information sur les gènes qui ne sont pas utiles pour l'étude des interactions entre gènes. Il est donc nécessaire de mettre au point des modules de prétraitement permettant d'extraire les zones textuelles pertinentes et de les restructurer le cas échéant. Ces opérations sont effectuées au moyen de programmes Perl et se composent de quatre étapes principales :

1. Sélection des zones textuelles pertinentes. Seuls le résumé qui suit le nom du gène et les sections `phenotypic infos` ainsi que `other infos` sont retenus car ce sont les seules sections à mentionner des interactions. Les sections `phenotypic infos` ainsi que `other infos` sont en fait constituées de brefs résumés d'articles scientifiques (7 à 8 phrases en moyenne), chacun précédé d'une référence à l'article en question. Pillet (2000) ne retient pas les résumés qui suivent le nom du gène dans les traitements qu'elle propose. Il s'agit pourtant d'une section importante dans la mesure où il existe des cas où une interaction est mentionnée dans le résumé et n'est pas reprise par la suite (cas du gène *14-3-3zeta* qui interagit avec *Src42A* sans que cette information soit reprise dans le corps de la notice).

2. Découpage en phrases. L'analyse du style des notices montre en effet que les interactions sont en général décrites au sein de la phrase, qui constitue le niveau optimal pour les traitements. Des erreurs interviennent toutefois assez fréquemment lors du découpage en phrases (titres non ponctués, retours à la ligne intempestifs, etc.).

3. Application d'heuristiques simples pour la résolution d'anaphores : le pronom impersonnel `it` en début de phrase est remplacé par le nom du gène décrit par la notice. Il est évident qu'une telle règle ne saurait s'appliquer telle quelle à d'autres corpus, mais elle fonctionne bien dans le cas de Flybase.

4. Le découpage en phrases pouvant séparer la phrase de sa référence (quand plusieurs phrases décrivent une même référence), il peut être nécessaire de l'ajouter en début de phrase par une procédure automatique.

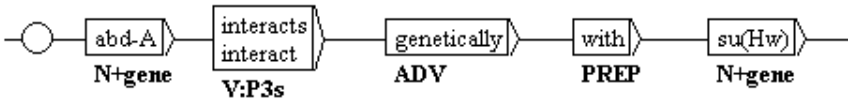
4 Repérage des noms de gènes

Les noms de gènes dans Flybase sont normalisés : tous les gènes sont désignés par leur symbole court. Ce n'est pas obligatoirement le cas d'autres bases de données comme Medline où, suivant l'espèce, les noms sont normalisés ou non. Des techniques d'acquisition et d'apprentissage s'avèrent alors intéressantes pour automatiser partiellement la procédure d'identification des noms de gènes (Proux *et al.*, 1998).

Pour Flybase, il n'existe pas, à notre connaissance, de liste des symboles directement accessible. On trouve en revanche une liste des noms longs et des symboles associés[1] qui, une fois nettoyée, permet d'obtenir un dictionnaire de plus de 23.000 noms de gènes. Comme ce dictionnaire est destiné à être utilisé avec Intex, on lui donne le format attendu par ce logiciel en concaténant à la fin de chaque entrée la catégorie du discours (`N` pour nom) et un trait sémantique :

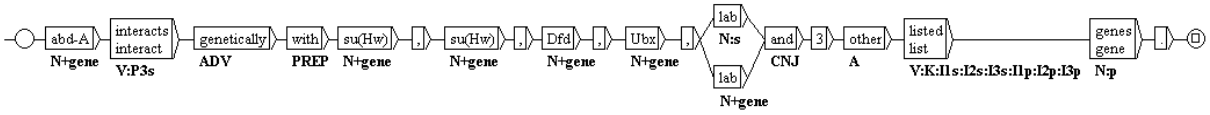
```
abd-A,.N+gene
Abd-B,.N+gene
lab,.N+gene
(…)
```

Le dictionnaire ainsi obtenu est intégré à Intex. Il permet un étiquetage du texte, où les noms de gènes sont reconnus en tant que tels. On obtient par exemple le fragment de graphe suivant :



Chaque mot a reçu une étiquette, qui figure sous la boîte du mot en question (cet exemple ne comporte pas de mot dont la partie du discours est ambiguë). Le jeu d'étiquette est relativement lisible (`N` pour nom, `V` pour verbe, `PREP` pour préposition, etc.) Certaines étiquettes sont assorties d'indications morpho-syntaxiques (par exemple, `V:P3s` désigne un verbe à la 3ème personne du présent de l'indicatif) ou de traits sémantiques (comme dans notre cas `N+gene` pour indiquer qu'un nom est un gène) (Silberztein 1993). Si le texte est ambigu, le système duplique certaines boîtes et leur affecte les étiquettes possibles.
Il n'est pas nécessaire de lever toutes les ambiguïtés des textes analysés (plus de 50 %. Il est toutefois nécessaire de désambiguïser au maximum les séquences où une étiquette `N+gene` est en concurrence avec une autre étiquette dans la mesure où c'est le nom de gène lui-même qui est alors ambigu. Dans l'exemple ci-dessous, le système n'a pas pu déterminer si le mot anglais *lab* désigne un laboratoire ou un nom de gène.

[1] Voir : `http://iubio.bio.indiana.edu:8099/fly`, lien `gene list`.



Dans des cas particuliers comme celui-ci, la structure de la séquence constitue un indice permettant la désambiguïsation. Les énumérations sont facilement repérables par l'accumulation de séquences étiquetées, séparées par des virgules ou la conjonction *and* (cf. (Daille *et al.*, 1996) pour une utilisation des énumérations dans un cadre d'analyse terminologique). Dans l'exemple ci-dessus, le nom de gène ambigu (*lab*) peut valablement être étiqueté en tant que gène, d'après le contexte (il figure dans une énumération de noms de gènes). Une dizaine de graphes (grammaires de désambiguïsation d'Intex) ont été développés pour agir à un niveau local et étiqueter correctement certaines séquences ambiguës. Certains noms de gènes rares et trop ambigus n'ont pas été pris e compte (*a*, *for*, *not*…).

5 Acquisition d'une liste de marqueurs d'interaction entre gènes

Une fois que les sections pertinentes ont été filtrées puis structurées et que les noms de gènes ont été mis en évidence, il importe de se pencher sur le repérage des interactions entre gènes. Si l'on considère les gènes comme des termes du domaine, la recherche s'oriente alors vers la mise au jour de marqueurs linguistiques traduisant une relation entre termes, autrement dit entre les gènes précédemment identifiés (Morin 1998, 1999). Cette relation est le plus souvent exprimée par un verbe (e.g *to interact*), plus rarement par une nominalisation (e.g. *interaction between…*). Elle est quasiment toujours explicite. On voit, à travers les textes de génomique, l'importance de l'analyse du verbe pour le repérage de relations entre entités puis l'élaboration d'un modèle conceptuel des interactions.

Il est donc nécessaire de repérer les marqueurs porteurs de la relation entre les deux gènes. La liste de marqueurs est établie en utilisant la structure particulière de Flybase : on sait que le résumé reprend en condensé l'information du corps de la notice. Il faut alors extraire du corpus les phrases où les mentions d'interaction mentionnés dans le résumé sont reprise en se fondant sur les phrases où l'on retrouve les noms de gènes mentionnés dans le résumé (à partir de la phrase stéréotypée `It interacts genetically with…`, où le verbe *interact* figure comme un hyperonyme des différents marqueurs d'interaction comme *activate* ou *inhibit*). On extrait de ces phrases les verbes apparaissant entre deux noms de gène, ce qui permet d'établir une liste d'une cinquantaine de « candidats marqueurs ». Une vingtaine de marqueurs d'interaction sont isolés, hors nominalisation (*interact*, *activate*, *inhibit*, *modulate*, *suppress*, *isolate*, *characterize, act, direct, repress, autoregulate, exert*…)[2]. A noter que l'on retrouve ainsi de manière semi-automatique la liste des verbes établie manuellement par (Thomas *et al.*, 2000) qui sert de support à leur étude (il s'agit de sept premiers verbes mentionnés dans la liste).

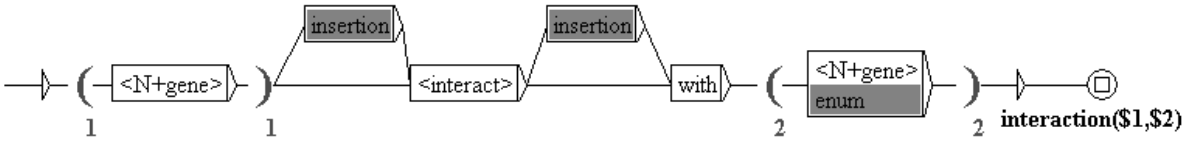
Cette stratégie d'amorçage est liée à la structure de Flybase et n'est pas reproductible telle quelle pour d'autres bases. Des expériences sur Medline montrent toutefois que d'autres

[2] Le repérage et l'acquisition des nominalisations est étudié de manière parallèle à travers les phénomènes de conservation partielle de la structure argumentale, pouvant être décrits au moyen de lexiques-grammaire (Balvet, 2001).

régularités permettant une acquisition semi-automatique existent, même si elles sont moins criantes. L'information dans Medline étant redondante, les stratégies fondées sur l'amorçage fonctionnent bien. Quelques interactions peuvent par exemple être acquises à partir d'une première base marqueurs. En focalisant ensuite la recherche sur les phrases où se trouvent présent les noms de gènes déjà repérés comme étant en interaction, de nouveaux marqueurs peuvent à leur tour être extraits : on atteint ainsi progressivement une bonne couverture du corpus (Morin, 1998). D'autres techniques d'apprentissage sont par ailleurs étudiées dans le cadre du projet Caderige[3] pour acquérir de manière automatique ou semi-automatique, non seulement des listes de verbes d'interaction, mais aussi leurs cadres de sous-catégorisation syntaxico-sémantique (Nedellec et Nazarenko, 2001).

6 Modélisation des transducteurs de repérage d'interaction

Des transducteurs reprenant les éléments essentiels de la relation au niveau linguistique (les deux noms de gènes et l'élément porteur de la relation) et permettant de généraliser le résultat obtenu en autorisant des insertions sont ensuite modélisés. Le transducteur suivant permet de reconnaître l'expression d'une interaction entre deux gènes :



Les boîtes en grisé sont en fait des appels à des sous graphes. Le sous-graphe `insertion` permet de gérer les cas d'insertion avant ou après le verbe (on autorise des insertions allant jusqu'à 5 mots, ces groupes de mots n'étant pas analysés syntaxiquement), le sous-graphe `enum` gère les énumérations comme dans la phrase `It interacts genetically with` ***`su(Hw) , su(Hw) , Dfd , Ubx , lab and 3 other listed genes.`*** Le graphe permet donc de reconnaître les séquences suivantes :

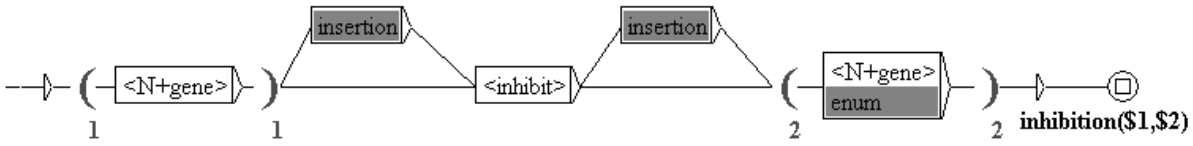
```
Abd-A interacts with su(Hw).
Abd-A interacts genetically  with su(Hw).
Abd-A interacts genetically  with su(Hw) , su(Hw) , Dfd , Ubx , lab
and 3 other listed genes.
```

Un transducteur possède en outre une fonction de réécriture qui permet d'extraire des fragments de séquences pour les réécrire sous une autre forme. On peut ainsi s'abstraire d'une séquence linguistique donnée pour la transformer en un équivalent de forme logique. Les fragments que l'on souhaite extraire doivent être entourés de parenthèses et sont repris dans les variables `$1`, `$2`, etc. Une forme logique unique est calculée, qui se présente sous la forme suivante :

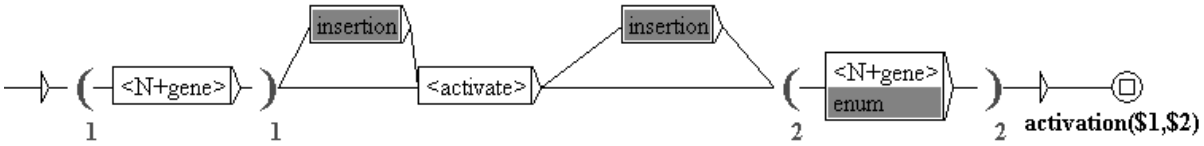
```
Interaction(Abd-A, su(Hw))
```

[3] Le projet Caderige a été labellisé dans le cadre de l'Action Bioinformatique CNRS-INRA-INRIA-INSERM. Il est composé de laboratoires de recherche en informatique et en biologie (projet AÏDA de l'IRISA, laboratoires CLIPS-IMAG, LIPN (U. de Villetaneuse), LRI (U. d'Orsay), laboratoires MIG, GM et LRV de l'INRA).

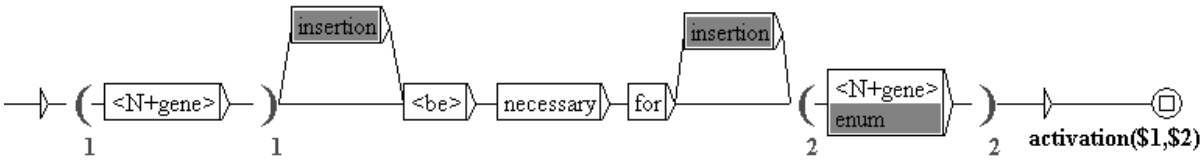
Un post-traitement est nécessaire pour décomposer les phrases où apparaît une énumération et générer autant de formes logiques que nécessaire. La relation entre les deux gènes est typée et orientée au moyen d'une famille de transducteurs d'extraction. On peut ainsi repérer des relations d'inhibition :



ou d'activation :



La forme logique calculée permet de regrouper des relations exprimées par des séquences linguistiques très différentes. La sortie du transducteur suivant génère une relation d'activation, même si la structure linguistique en cours d'analyse ne contient ni *activate* ni un de ses dérivés :



Ce transducteur montre que les relations exprimées par le verbe *activate* et l'expression *to be necessary for* induisent une relation du même type (*activation*). Ces transducteurs sont pour l'instant élaborés à la main, mais les partenaires du projet Caderige ont pour ambition d'utiliser des techniques d'apprentissage pour automatiser au moins partiellement leur acquisition à partir du corpus.

La forme logique peut être assimilée à un niveau conceptuel, où les éléments sont normalisés indépendamment de la séquence linguistique effectivement présente dans les textes. Il est ainsi possible de constituer une base de connaissances sous la forme de prédicats Prolog pouvant être directement interrogés.

7 Évaluation

Les différentes étapes du processus ont été évaluées. Pour ce faire, on a isolé un corpus de cent notices distinctes de celles sur lesquelles le système a été mis au point. Ces cent notices forment un corpus de 1,9 Méga-octets. La précision correspond à la proportion d'interactions effectives par rapport au nombre d'interaction relevé. Le rappel correspond à la proportion d'interactions relevées par rapport à celles qui auraient théoriquement dû être repérées.

La recherche des noms de gènes dans Flybase ne pose pas de problèmes majeurs, dans la mesure où les noms sont normalisés (Flybase utilise les noms de symbole courts). La

précision est quasiment parfaite dans la mesure où l'on n'étiquette que des noms de gènes pas ou peu ambigus (les assurances que l'on prend pour l'étiquetage des séquences ambiguës sont suffisamment fortes pour ne pas laisser passer de bruit). La prudence dont on a fait preuve s'est en revanche traduite par du silence, mais celui-ci reste faible, d 'après une évaluation manuelle (inférieur à 2 %). En effet, les noms de gènes très ambigus éliminés du dictionnaire n'apparaissent quasiment jamais dans des séquences d'interaction[4].

Un ensemble de transducteurs de mise en relation a été défini manuellement, d'après la liste de verbes acquis à partir du corpus (section 5). Le système permet d'extraire 137 instances de relation du corpus, dont 58 seulement sont différentes[5] (sur ces 58 interactions, 37 sont des activations et 5 des inhibitions), du fait que la plupart des interactions décrites dans le résumé sont reprises dans le corps du texte. La précision est bonne et se situe au-dessus de 95 % (d'après un relevé fait à la main, avec le texte en regard des résultats[6]). Ce bon résultat est dû au fait que l'information recherchée est bien bornée : on recherche des interactions entre gènes à partir de textes où les gènes sont étiquetés de manière non ambiguë et où les marqueurs de relations sont précis et validés manuellement. Le pourcentage d'erreur (ici moins de 5 %) est dû aux insertions : les insertions sont décrites au moyen de graphes peu déterminés : une insertion est une séquence de mots d'une longueur donnée. La sous-détermination de la description des insertions n'est pas un problème majeur dans la mesure où elle est contrebalancée par la précision des autres éléments décrits (marqueurs de relations et noms de gènes dont la présence est obligatoire). Dans le cas présent, on a limité les insertions à 5 mots ; le taux d'erreur serait plus fort si on augmentait la taille des insertions possibles.

Le rappel (complémentaire du taux de silence) est plus difficile à déterminer. On peut distinguer deux cas :

1. une séquence n'a pas été extraite alors que tous les éléments (noms de gènes, marqueur de relation) avaient été modélisés dans le système,

2. le silence est dû à un élément non modélisé (nom de gène éliminé car trop ambigu, marqueur absent…).

Le premier cas se produit quand la structure linguistique est trop complexe et a échappé à l'analyse. Vu que, dans ce cas, tous les éléments nécessaires à l'identification ont été modélisés, c'est souvent la nature des insertions qui est trop complexe. Les cas suivants ont pu être relevés suite à un examen manuel : insertion trop longue (on a pris un cadre assez

---

[4] Les noms de gènes *a*, *for*, *while*, ne sont jamais mentionnés comme acteurs d'une interaction (Pillet, 2000).

[5] L'étude de Pillet (2000), bien que portant sur toute la base (soit plus de 20.000 notices), ne permet d'extraire que 653 phrases où il est question d'interaction. Ce chiffre, divergent par rapport au nôtre, est probablement dû au fait que les traitements proposés par l'auteur ne portent que sur la section « Phenotypic info. » de la base et que seules les phrases où deux noms de gènes apparaissent sont retenues, alors que les énumérations ne sont pas rares.

[6] Pillet (2000) obtient un taux de précision de 87 %, mais sur une tâche plus simple (catégorisation des phrases suivant qu'elles mentionnent ou non une interaction). Ce taux est à relativiser dans la mesure où le corpus de test de l'étude de Pillet est le même que le corpus d'entraînement. Les taux couramment mentionnés dans les études fondées sur des méthodes statistiques vont plutôt de 60 % à 80 % de précision.

restrictif de 5 mots), présence d'un élément non reconnu (il suffit d'un élément comme une ponctuation non modélisée dans l'automate pour que la reconnaissance échoue), etc. Il serait possible d'améliorer les choses en passant sur les textes des automates de simplification de texte (automates éliminant certaines structures comme les insertions ou les relatives, permettant de travailler sur un texte épuré).

Le deuxième cas est le plus fréquent, la couverture du système reste insuffisante. Dans cette expérience, l'accent a été mis sur la mise en place d'une méthode et des outils pour le repérage des interactions avec une très grande précision, plus que sur la couverture du système[7]. Un effort pour repérer puis modéliser de nouveaux marqueurs reste nécessaire. De ce point de vue, un corpus d'apprentissage plus large serait utile (on s'est limité ici volontairement à cent notices). Le traitement des nominalisations et des anaphores permettrait de mieux couvrir le corpus (on estime qu'environ 1/5^{e} des relations sont mentionnées sous forme de nominalisations).

8 Conclusion et perspectives

Cet article a présenté une méthodologie et des outils pour le repérage d'interactions entre gènes au moyen de transducteurs à nombre fini d'états. Les traitements permettent de passer de grandes bases textuelles en génomique vers des tables structurées d'interactions entre gènes. Les outils proposés permettent en outre une normalisation des relations qui sont typées et orientées. Il est ainsi possible de s'abstraire du niveau linguistique pour passer à un niveau conceptuel, à travers une forme logique qui normalise les différentes séquences textuelles apparaissant en corpus. Les résultats obtenus permettent d'obtenir une très bonne précision avec un rappel plus médiocre.

L'effort doit donc être porté sur l'élaboration de bases de marqueurs de relation beaucoup plus importantes. Pour cela, il est possible de recourir à des techniques d'acquisition et d'apprentissage. C'est le but de projets en cours comme le projet Caderige qui vise à acquérir de manière automatique ou semi-automatique des ressources lexicales, syntaxiques et sémantiques pour l'extraction de connaissances à partir de bases de données textuelles en génomique (Nedellec et Nazarenko, 2001). Caderige s'intéressera par ailleurs à Medline, qui constitue un terrain d'expérimentation à la fois plus riche, mais aussi plus difficile que Flybase.

Remerciement

Le contenu de cet article doit beaucoup à des discussions que j'ai pu avoir les partenaires du projet Caderige, en particulier Adeline Nazarenko (LIPN) et Claire Nedellec (LRI). Je les remercie pour leur relecture minutieuse qui m'a permis d'améliorer bien des points de cet article.

---

[7] L'article de (Thomas *et al*, 2000) va dans le même sens, sachant que sur de grandes bases de données, la redondance des informations fait qu'une interaction entre gènes est généralement mentionnée plusieurs fois. Il est donc préférable à leurs yeux de mettre en avant la précision plutôt que le rappel.

Références

Balvet, A. (2001). « Grammaires locales et lexique-grammaire pour le filtrage par *chunks* ». *Actes de Terminologie et Intelligence Artificielle* (TIA'2001), Nancy.

Daille, B., Habert, B., Jacquemin, C., Royauté, J. (1996). « Empirical observation of term variations and principles for their description ». *Terminology*, vol. 3(2), pp. 197–258.

Flybase. US National Institute of Health and British Medical Research. (disponible à l'adresse : `http://flybase.bio.indiana.edu/`)

Goujon, B. (1999). « Extraction d'informations techniques pour la veille par l'exploitation de notions indépendantes d'un domaine ». *Terminologies Nouvelles*, vol. 19, pp. 33–42.

Humphreys, K., Demetriou, G., Gaizauskas, R. (2000). « Two applications of information extraction to biological science article: enzyme interaction and protein structures ». *Proceedings of the Pacific Symposium on biocomputing* (PSB'2000), Honolulu, vol. 5, pp. 502–513.

DARPA (1995). *Proceedings of the Sixth Message Understanding Conference*. Morgan Kaufmann Publishers, San Francisco.

Morin, E. (1998). « Prométhée : un outil d'aide à l'acquisition de relations entre termes », *Actes Traitement Automatique des Langues Naturelles* (TALN'1998), Paris, pp. 172–181.

Morin E. (1999). Extraction de liens sémantiques entre termes à partir de corpus de textes techniques. *Thèse de l'Université de Nantes*.

Nedellec, C., Nazarenko, A. (2001). « Application de l'apprentissage à la recherche et à l'extraction d'information - Un exemple, le projet *Caderige* : identification d'interactions géniques ». *Journée sur l'exploration de données issues d'Internet*, Université de Villetaneuse.

Pillet, V. (2000). Méthodologie d'extraction automatique d'information à partir de la littérature scientifique en vue d'alimenter un nouveau système d'extraction. *Thèse de l'Université d'Aix-Marseille 3*.

Proux, D., Rechenmann, D., Julliard, L., Pillet, V. Jacq, B. (1998). « Detecting gene names in biological texts: toward information extraction about molecular interactions ». *Proceeding of the ninth workshop on genome informatics*, Tokyo, pp. 72–80.

Silberztein, M. (1993). *Dictionnaires électroniques et analyse automatique des langues : le système Intex*. Masson, Paris.

Stapley, B.J., Benoit, G. (2000). « Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts ». *Proceedings of the Pacific Symposium on biocomputing* (PSB'2000), Honolulu, vol. 5, pp. 491–502.

Thomas, J., Milward, D., Ouzounis, C., Pulman, S., Carroll, M. (2000). « Automatic extraction of protein interactions from scientific abstracts ». *Proceedings of the Pacific Symposium on biocomputing* (PSB'2000), Honolulu, vol. 5, pp. 538–549.