Analyse de la littérature biomédicale pour détecter les interactions entre protéines: comparaison du Text Mining sur les résumés et sur les articles en texte intégral.

Eric Martin[1] et Olivier Jouve[1]

{emartin, ojouve}@spss.com
[1]SPSS, Tour Europlaza, La Défense 4, F-92925 Paris-la-Défense Cedex, France

Mots clefs :
Traitement du Langage Naturel, Text Mining, Data Mining, Extraction d'Informations d'Articles Biomédicaux, Protéomique, Bioinformatique, Bibliométrie, Veille Technologique, Gestion des Connaissances.

Keywords:
Natural Language Processing, Text Mining, Data Mining, Biomedical Literature Mining, Proteomics, Bioinformatics, Bibliometrics, Scientific Watch, Knowledge Management

Palabras clave :
Tratamiento en Lenguaje Natural, Minería de Texto, Minería de Datos, Extracción de información de Artículos Biomédicos, Proteómica, Bioinformática, Bibliometría, Vigilancia Tecnológica, Gestión del Conocimiento.

Résumé.

Le problème de la gestion des connaissances dans l'industrie pharmaceutique est double. Il s'agit tout d'abord de pouvoir intégrer les données provenant du séquençage du Génome avec les données provenant l'analyse fonctionnelle des gènes. Parallèlement, alors que le nombre de publications biomédicales croît de manière exponentielle (Medline contient de nos jours plus de 14 millions d'enregistrements), les chercheurs ont besoin d'être assistés pour élargir et approfondir la compréhension des nombreux domaines scientifiques. En se focalisant sur l'analyse de textes libres, le text mining permet de structurer les connaissances, d'en définir les contours et de découvrir de nouvelles tendances. Nous décrivons dans le présent article une méthode pour détecter automatiquement les interactions entre protéines à partir d'un grand nombre de publications. Cette méthode utilise l'analyse en langage naturel pour identifier le nom des protéines, leurs synonymes ainsi que les diverses interactions qu'elles peuvent présenter entre elles. Nous avons par la suite comparé l'analyse en text mining sur les résumés (résumés) à celle réalisée sur les articles complets (full text) de façon à déterminer la quantité de relations perdues lorsque seuls les résumés sont exploités. Nos résultats montrent que : 1) LexiQuest Mine se révèle être un outil extrêmement précis et polyvalent pour l'analyse des interactions entres protéines à partir des articles biomédicaux. 2)L'analyse des résumés peut être suffisante et faire gagner un temps précieux pour des applications ne nécessitant pas un niveau de détail très élevé alors que l'analyse des articles complets est de rigueur pour des applications très ciblées.

1 Introduction

Il existe une grande différence entre information et connaissance. Le volume d'informations auquel l'industrie pharmaceutique doit faire face a explosé depuis quelques années. Avec l'achèvement du projet Human Genome, une grande quantité d'informations est à présent disponible sur les gènes humains. La connaissance de la manière dont ces gènes fonctionnent et tout particulièrement dont les protéines interagissent entre elles pour former de complexes cycles biochimiques est très difficile à appréhender et à maîtriser. Cette connaissance est pourtant capitale aux chercheurs pour développer de nouveaux médicaments traitant la maladie à sa source.

De nombreuses bases de données ont été crées pour fournir des informations sur les protéines et leurs interactions. Parallèlement, des initiatives de création d'ontologies telles que Gene Ontology (GO) permettent de faire progresser l'analyse *in silico* des résultats scientifiques. Cependant la littérature scientifique reste la source la plus riche de résultats expérimentaux et cliniques. Le nombre de publications scientifiques dans le domaine des sciences du vivant ne cesse d'augmenter : Medline [1] comprend plus de 14 millions d'articles et croît d'au moins 4% par an. Face à une telle explosion de données et une telle complexité, les chercheurs ont un réel besoin d'être assistés d'outils informatiques pour élargir et approfondir leur compréhension des interactions entre protéines.

Les résumés ainsi que les articles complets (full text) sont utilisés depuis longtemps par la communauté scientifique. Les résumés sont facilement accessibles mais sont d'un contenu limité tandis que les articles en texte intégral fournissent un contenu beaucoup plus riche mais sont plus délicats à analyser car ils sont plus difficile d'accès, demandent des capacités de traitement bien plus grandes et présentent un haut niveau d'hétérogénéité de formats et de présentations.

L'analyse des résumés ou des documents en texte intégral nécessite tout d'abord l'identification des entités (dans notre cas les protéines) considérées puis la compréhension des relations complexes qui lient ces entités. Les techniques utilisées pour l'identification d'entités biologiques (protéines, gènes, molécules,…) incluent des méthodes à base de règles [2, 3], à base de dictionnaires [4-6], ainsi que des méthodes à partir d'algorithmes d'apprentissage [7-10]. Le processus complexe d'analyse des relations est aussi un domaine très en vogue ; cela comprend l'analyse des interactions entre protéines [11-16], l'annotation de gènes [17], la compréhension des relations entre les gènes et les molécules thérapeutiques [18] et l'identification des voies métaboliques [19]. Les techniques varient de l'utilisation de règles basées sur des patrons [11, 12, 17], à des techniques de « shallow parsing » [13, 14, 16] en passant par l'analyse grammaticale des phrases (« full sentence parsers) [16]. Les systèmes à base de règles utilisent des motifs pré-définis (comme l'existence de mots spécifiques à des endroits spécifiques de la phrase). Les « shallow parsers » décomposent partiellement les phrases pour identifier certains composants et leurs dépendances locales sans considérer l'ensemble de la phrase, contrairement aux « full sentence parsers ». La précision et le rappel de tels systèmes ont été décrits comme allant de 60% à 95%. Cependant dans la plupart des cas, ces mesures étaient basées sur l'analyse d'un seul type d'interaction [15, 17] ou sur un très petit corpus [11, 12, 17]. Par exemple, Sekimizu *et al* [17] ont seulement extrait les relations associées à sept verbes fréquents (activate, bind, interact, regulate encode, signal, and function) trouvés dans les résumés de Medline. La précision et le rappel de ces systèmes varie de 67.8% à 83.3% selon le verbe utilisé. Ono *et al.* [11] ont mesuré une précision de 94% et un rappel de 83% pour des textes traitant uniquement du génome de la levure. Pustejovsky *et al.* [17] quant à eux font état d'une précision de 90% and d'un rappel de 57% à partir d'un corpus de 500 abstrats annotés manuellement pour un seul type d'interaction.

Nous décrivons ici une méthode de text mining pour détecter automatiquement des interactions entre protéines à partir de corpus volumineux. Cette méthode est basée sur l'analyse en langage naturel des noms des protéines, de leurs synonymes ainsi que des multiples interactions qu'elles peuvent présenter avec d'autres protéines. Nous avons aussi voulu examiner les avantages et inconvénients du text mining sur des résumés comparés à celui des articles complets. Nous avons ainsi analysé 659 résumés d'articles scientifiques et les avons comparé aux articles intégraux correspondant. L'analyse des résultats nous a permis de mesurer la précision et le rappel de LexiQuest Mine quant à l'analyse sémantique des interactions entre protéines à partir des résumés ou des articles complets.

2 Matériel et Méthodes

2.1 Construction du corpus

La base de données DIP™ [20] database recense les résultats expérimentaux d'interactions entre protéines. Elle agrège des informations de multiples sources pour fournir un point d'accès unique aux interactions entre protéines. Une requête sur DIP a permis d'identifier les articles possédant au moins une interaction recensée entre deux protéines. 659 articles furent choisis et récupérés sous la forme d résumés et de textes intégraux au format pdf pour constituer le corpus final.

Ce corpus a été constitué de façon à ce que le contenu soit hétérogène et représentatif de Medline : on y trouve des interactions entre protéines dans différents systèmes expérimentaux comme la drosophile, la levure, la souris ou bien encore des lignées cellulaires humaines.

2.2 LexiQuest Mine 2.2

LexiQuest Mine [21] est un logiciel commercial de text mining développé et distribué par SPSS. Il est basé sur des technologies linguistiques d'analyse en langage naturel (NLP, Natural Language Processing). LexiQuest Mine identifie automatiquement les concepts clés ainsi que leur nature (nom de personne, d'organisation, de lieu, de protéine, de molécule, …) puis les relations sémantiques qui les lient. LexiQuest Mine possède deux modes opératoires distincts qui peuvent être utilisés séparément ou ensemble:

- Extraction des concepts, basée sur l'analyse de syntagmes nominaux
- Analyse des relations sémantiques, basée sur l'analyse de patrons linguistiques dynamiques

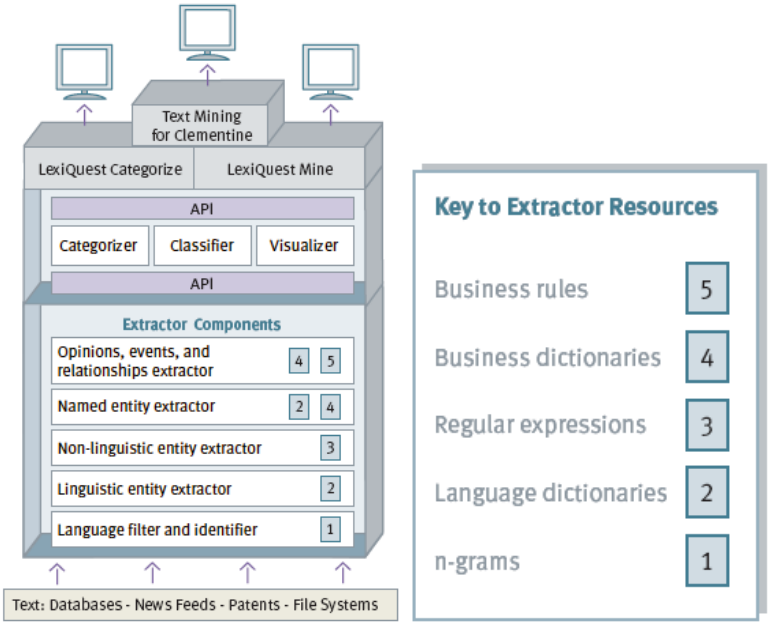

Figure1. Schéma de la solution de Text Mining; le même moteur linguistique est utilisé par trois différents produits: LexiQuest Mine [21], LexiQuest Categorize [22] et Text Mining for Clementine [23].

Pour analyser les interactions entre protéines, LexiQuest Mine utilise :

- Des dictionnaires de noms de protéines (i.e. il-10, nf-kappa b, …).
- Des dictionnaires de synonymes (Il-10=interleukin 10; NF-kappaB=nuclear factor kappab)
- Des listes de marqueurs linguistiques pour détecter des protéines inconnues.
- Des dictionnaires d'interactions (binds, inhibits, activates, …)
- Des listes d'expressions régulières

- Des patrons dynamiques, comme celui décrit fig.2. Plusieurs dizaines de patrons ont été utilisés lors de cette étude.

Tous les dictionnaires et fichiers de ressources peuvent être modifiés par les utilisateurs.

```
sentence: Exogenous IL-10 inhibits NF-kappaB in monocytes.
pattern(300)]
name = (300)_G_AV_G_1
value = ($VarGene $Verb (the)? $VarGene)
output = interleukin 10 inhibits nf-kappa b
```

Figure 2. Exemple de patrons linguistique utilisé pour l'identification des interactions entre protéines: la phrase analysée correspond au patron n°300 dans lequel deux gènes ou protéines sont séparés par un verbe à la forme active. La sortie permet de comprendre que : Interleukin 10 inhibits nf-kappa b.

3 Résultats

3.1 Optimisation des phases de Text Mining

Le but premier de cette étude est de montrer la faisabilité de l'analyse par LexiQuest Mine des interactions protéine/protéine décrites dans les publications scientifiques. Deux thesaurri ont été utilisés pour identifier les protéines : GeneOntology et celui fournit par défaut avec LexiQuest Mine pour les applications de génomique. De plus, des phases préliminaires d'extraction de concepts et de typage automatique des protéines ont permis d'améliorer sensiblement la qualité des dictionnaires.

Tous les principaux types d'interactions protéine/protéine ont tout d'abord été listés par des experts puis les patrons linguistiques ont été adaptés de façon à reconnaître ces différentes interactions dans des contextes très divers. Un exemple des relations étudiées est fourni dans le tableau 1.

Tableau 1. Liste des principaux prédicats (à la forme active) utilisés pour analyser les interactions entre protéines

Activates	Forms complex with	Over-expresses
Antagonizes	Inactivates	Reduces
Associates with	Increases	Regulates
Binds	Links	Releases
Down-regulates	Mediates	Represses

Les patrons utilisés sont capables de prendre en compte des relations complexes entre 2 à 5 protéines ainsi que de faire la différence entre les formes affirmatives et négatives des relations étudiées.

Les résultats partiels provenant de l'analyse des patrons (pattern matching) ont aussi été analysés afin de trouver de nouvelles protéines ou interactions : LexiQuest Mine fourni des résultats à 4 colonnes comprenant l'identifiant du document et les deux entités nommées séparées par la relation qui les lie. Les résultats sont listés dans le tableau2.

Tableau 2. Types de résultats partiels

1.	Protein A		Protein B
2.		Interacts with	Protein B
3.	Protein A	Interacts with	

D'après le tableau2, il apparaît clairement que les cas 2 et 3 sont intéressants pour détecter de nouveaux noms de protéines qui n'auraient pas été préalablement identifiés tandis que le cas 1 permet de détecter de nouveaux types d'interactions. Par exemple, le cas 1 permet de trouver des phrases du type:

*together with our finding of coimmunoprecipitation of <*act3p*> with histone <*h2a*> , this suggests the in vivo existence of a protein complex required for correct expression of particular genes*.

Ainsi, "coimmunoprecipiation" a pu facilement être ajouté à la liste des interactions possibles entre protéines.

3.2 Détection automatique des interactions entre protéines

Les articles biomédicaux ayant été identifiés à partir de DIP, chaque document full text est censé contenir au moins une interaction entre au moins deux protéines. Après analyse par text mining, seuls les résumés présentant au moins une interaction valide furent retenus, ainsi que les documents full text correspondant. La proportion de documents positifs après analyse est décrite dans le tableau 3. Les résultats montrent que moins de 20% des articles complets furent ignorés par LexiQuest Mine. Bien que représentant moins de 3.3% de la taille du corpus full text, plus de 53% des résumés furent détectés comme positifs.

Tableau 3: Proportion de documents détectés comme positifs (possédant des interactions entre protéines).

	Nombre total de documents	**Proportion de documents positifs**
Résumés seuls	659	53.41%
Full text	659	81.49%

Les documents positifs furent gardés pour les analyses ultérieures. L'analyse quantitative des relations est donnée par le tableau 4. La grande proportion de relations uniques reflète la grande hétérogénéité du corpus traité. Les résumés ne contiennent que peu d'interactions entre protéines car ils ont pour but de synthétiser les résultats majeurs. A l'opposé, les documents full text contiennent beaucoup plus d'interactions, mais leur nombre varie énormément d'un article à l'autre.

Tableau 4. Statistiques sur les documents positifs. Une relation unique correspond à une séquence unique: PMID - protéine A - relation- protéine B

	Nombre total de relations identifiées	**Proportion de relations uniques**	**Nbre moyen de relations par doc.**	**Ecart type du nbre de relations par doc.**
Résumés seuls	853	92.26%	2.42	1.65
Full text	8098	60.21%	15.08	14.11

La précision et le rappel constituent deux mesures de choix pour l'évaluation des moteurs de recherche. Ils peuvent être ici aussi utilisés de la manière suivante:

$$\Pr écision = \frac{\mathrm{Re}\, lations\_exactes}{(\mathrm{Re}\, lations\_exactes + \mathrm{Re}\, lations\_inexactes)}$$

$$Rappel = \frac{\mathrm{Re}\, lations\_exactes}{(\mathrm{Re}\, lations\_exactes + \mathrm{Re}\, lations\_manquées)}$$

Evaluer la précision et le rappel demande un énorme travail manuel pour identifier l'ensemble des relations manquées et celui des relations incorrectes. Un échantillon de 32 documents (résumés et full-text) a donc été sélectionné pour ce faire. Les mêmes mesures que pour le tableau 4 ont été réalisées dans le tableau 5 de manière à évaluer la qualité de l'échantillon.

Tableau 5. Statistiques sur l'échantillon composé des 32 documents

	Nre de relations	Proportion de relations uniques	Nbre moyen de relations par doc.	Ecart type du nbre de relations par doc.
Résumés seuls	91	81.32%	2.75	2.45
Full text	796	59.80%	24.37	16.88

L'échantillon montre des propriétés proches du corpus initial et a été considéré comme statistiquement représentatif. Chaque document a été analysé manuellement par des spécialistes et leurs résultats ont été comparés à ceux de LexiQuest Mine pour mesurer la précision et le rappel du système (Tableau 6).

Tableau 6. Analyse de la précision et du rappel sur l'échantillon de documents

	Nombre de documents	Précision	Rappel
Résumés échantillon	32	92.5%	66.1%
Full text échantillon	32	82.8%	64.1%

Les résultats montrent un niveau de précision sur les résumés remarquable pour un corpus hétérogène. Il est de plus intéressant de noter que malgré les problèmes de formats et de présentation des articles full text, le niveau de précision reste très élevé, montrant ainsi la robustesse des patrons linguistiques. Autour des 65%, le rappel est lui aussi plus qu'honorable, d'autant plus qu'il reste quasiment identique entre résumés et articles full text.

4 Discussion

4.1 Text Mining pour l'analyse des interactions entre protéines : limites et améliorations.

Les interactions non détectées par notre système peuvent avoir de multiples causes:

- Un haut niveau de complexité grammaticale
- Un problème lors de la conversion pdf->txt qui peut altérer parfois les phrases
- Un défaut des dictionnaires ou des algorithmes destinés à détecter les noms des protéines ou le type d'interaction en jeu.

Cependant le fait de travailler au niveau de la phrase permet au système d'être suffisamment redondant pour finalement trouver la majeure partie des interactions sans pour autant augmenter le bruit.

Comparées aux approches statistiques telles que les réseaux bayésiens, les approches linguistiques présentent de nombreux avantages :

- Le temps de traitement est beaucoup plus rapide et permet ainsi le déploiement de véritables applications professionnelles et transversales au sein des entreprises.
- Le niveau de précision est plus élevé : les interactions trouvées sont analysées sémantiquement au lieu de simplement être détectées comme statistiquement probables.
- La traçabilité du traitement : chaque résultat peut être clairement explicité.

Cependant, les technologies à base d'analyse en langage naturel nécessitent d'être méticuleux quant à la constitution des dictionnaires d'entités nommées [2, 24, 25, 26], de relations ou encore des patrons dynamiques pour l'analyse des relations. Même si des nomenclatures se mettent en place pour uniformiser les entités nommées, elles ne sont le plus souvent pas respectées par la communauté des biologistes qui utilise une grande variété de modèles expérimentaux et qui est composée de cultures scientifiques très variées [24].
Plusieurs autre systèmes similaires ont déjà été décrits pour l'analyse automatique des interactions entre protéines [11, 12, 27-32]. SUISEKI (System for Information Extraction on Interactions) par exemple présente une précision autour de 50% et un rappel de l'ordre de 20%. La meilleure qualité de nos résultats peut s'expliquer à différents niveaux :

- Un moteur d'analyse en langage naturel sophistiqué
- Un grand nombre de patrons dynamiques pour décrire les relations entre protéines, prenant par exemple en compte la coordination et la négation
- Une optimisation semi-automatique des listes d'entités nommées.

4.2 Résumé ou texte intégral ?

Traiter les articles complets permet évidemment de récupérer beaucoup plus d'interactions mais ne change que très peu la précision et le rappel de notre système. Dans ces conditions, l'analyse des résumés présente de nombreux avantages :

- Ils peuvent facilement être récupérés au format txt, évitant ainsi les problèmes de conversion entre différents formats.
- L'information a été triée : seuls les résultats notables sont généralement mentionnés.
- Les relations sont souvent exprimées de façon plus concise que dans les articles full text, rendant le text mining plus aisé.
- Ils sont accessibles gratuitement.
- Le temps de traitement est beaucoup plus court.

Il nous semble donc que les résumés constituent un matériel de choix pour les applications de text mining à grande échelle, ne nécessitant pas un niveau très élevé d'exhaustivité mais demandant l'analyse rapide d'un grand nombre de sources d'informations avec un haut niveau de précision.
D'un autre côté, l'analyse des articles full text nous semble préférable lorsque:

- L'accès à ces documents est aisé.
- Le nombre d'entités nommées et de relations étudiées est limité et bien défini.
- L'analyse des données et méthodes expérimentales est requise.
- Le text mining est orienté vers la veille technologique pour détecter de nouvelles tendances ou des relations encore peu connues par la majeure partie de la communauté scientifique.

5 Conclusion

Avec des résultats supérieurs à 80% de précision et 65% de rappel, notre système présente un faible taux d'interactions faussement positives, tout en garantissant une précision élevée, que ce soit sur les résumés ou sur l'intégralité des articles. La plupart des associations bien connues entre protéines ont été retrouvées à partir des résumés. D'un autre côté, l'analyse des documents full text permet de récupérer environ dix fois plus de résultats sans pour autant diminuer de façon drastique la précision.
Bien que cette étude se soit focalisée sur l'analyse des interactions entre protéines, LexiQuest Mine peut être utilisé pour l'analyse de tout type de relations entre tout type d'entités nommées.
LexiQuest Mine a déjà été déployé avec succès dans de nombreux laboratoires privés et universitaires. Ces résultats laissent à penser que LexiQuest Mine est un outil de choix lorsqu'il s'agit d'inclure des solutions de Text Mining dans les plateformes classiques de bioinformatique ou de veille technologique. La veille technologique s'en trouve profondément améliorée.

Références

1. National Library of Medicine's bibliographic database at http://www.ncbi.nlm.nih.gov
2. Fukuda, K., et al., *Toward information extraction: identifying protein names from biological papers.* Pac Symp Biocomput, 1998: p. 707-18.
3. Narayanaswamy, M., K.E. Ravikumar, and K. Vijay-Shanker, *A biological named entity recognizer.* Pac Symp Biocomput, 2003: p. 427-38.
4. Krauthammer, M., et al., *Using BLAST for identifying gene and protein names in journal articles.* Gene, 2000. **259**(1-2): p. 245-52.
5. Hanisch, D., et al., *Playing biology's name game: identifying protein names in scientific text.* Pac Symp Biocomput, 2003: p. 403-14.
6. Egorov, S., A. Yuryev, and N. Daraselia, *A simple and practical dictionary-based approach for identification of proteins in Medline abstracts.* J Am Med Inform Assoc, 2004. **11**(3): p. 174-8.
7. Hatzivassiloglou, V., P.A. Duboue, and A. Rzhetsky, *Disambiguating proteins, genes, and RNA in text: a machine learning approach.* Bioinformatics, 2001. **17 Suppl 1**: p. S97-106.
8. Wilbur, W.J., et al., *Analysis of biomedical text for chemical names: a comparison of three methods.* Proc AMIA Symp, 1999: p. 176-80.
9. Collier, N., C. Nobata, and T. Tsujii, *Extraction of name of genes and gene products with a Hidden Markov Model.* COLING conference proceedings, 2000.
10. Kazama, J., et al., *Tuning Support Vector Machines for Biomedical Named Entity Recognition.* Proceedings of the Natural Language Processing in the Biomedical Domain, 2002.
11. Ono, T., et al., *Automated extraction of information on protein-protein interactions from the biological literature.* Bioinformatics, 2001. **17**(2): p. 155-61.
12. Wong, L., *PIES, a protein interaction extraction system.* Pac Symp Biocomput, 2001: p. 520-31.
13. Humphreys, K., G. Demetriou, and R. Gaizauskas, *Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures.* Pac Symp Biocomput, 2000: p. 505-16.
14. Park, J.C., H.S. Kim, and J.J. Kim, *Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar.* Pac Symp Biocomput, 2001: p. 396-407.
15. Pustejovsky, J., et al., *Robust relational parsing over biomedical literature: extracting inhibit relations.* Pac Symp Biocomput, 2002: p. 362-73.
16. Yakushiji, A., et al., *Event extraction from biomedical papers using a full parser.* Pac Symp Biocomput, 2001: p. 408-19.
17. Sekimizu, T., H.S. Park, and J. Tsujii, *Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts.* Genome Inform Ser Workshop Genome Inform, 1998. **9**: p. 62-71.
18. Rindflesch, T.C., et al., *EDGAR: extraction of drugs, genes and relations from the biomedical literature.* Pac Symp Biocomput, 2000: p. 517-28.
19. Ng, S.K. and M. Wong, *Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts.* Genome Inform Ser Workshop Genome Inform, 1999. **10**: p. 104-112.
20. http://dip.doe-mbi.ucla.edu
21. http://www.spss.com/lexiquest/lexiquest_mine.htm
22. http://www.spss.com/lexiquest/lexiquest_categorize.htm
23. http://www.spss.com/lexiquest/text_mining_for_clementine.htm
24. Franzen, K., et al., *Protein names and how to find them.* Int J Med Inf, 2002. **67**(1-3): p. 49-61.
25. Tanabe, L. and W.J. Wilbur, *Tagging gene and protein names in biomedical text.* Bioinformatics, 2002. **18**(8): p. 1124-32.
26. Jouve, O., Harrison, L., Borgulia, P. and B. Simblist, *A lexical approach to text mining for the integration of genomic infomation*. Proceedings of the Atlantic Symposium on Computational Biology and Genome Information Systems and Technology, 2001.

27. Blaschke, C. and A. Valencia, *The potential use of SUISEKI as a protein interaction discovery tool*. Genome Inform Ser Workshop Genome Inform, 2001. **12**: p. 123-34.
28. Hu, X., et al., *Extracting and Mining Protein-Protein Interaction Network from Biomedical Literature.* Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2004.
29. Daraselia, N., et al., *Extracting human protein interactions from MEDLINE using a full-sentence parser*. Bioinformatics, 2004. **20**(5): p. 604-11.
30. Huang, M., et al., *Discovering patterns to extract protein-protein interactions from full texts.* Bioinformatics, 2004.
31. Marcotte, E.M., I. Xenarios, and D. Eisenberg, *Mining literature for protein-protein interactions.* Bioinformatics, 2001. **17**(4): p. 359-63.
32. Temkin, J.M. and M.R. Gilder, *Extraction of protein interaction information from unstructured text using a context-free grammar.* Bioinformatics, 2003. **19**(16): p. 2046-53.