Approche terminologique et infométrique de fouille de données textuelles dans un corpus sur la génétique des cancers de la thyroïde

A Terminological and Informetric Approach for Text Mining in a Corpus Involved in the Genetics of Thyroid Cancers

J. Royauté[1], C. François[2], A. Zasadzinski[2], D.Besagni[2], P. Dessen[3], S. Le Minor[3], MT. Maunoury[3]

[1] Laboratoire d'Informatique Fondamentale de Marseille (LIF), UMR 6166 CNRS.
[2]Unité de Recherche et Innovation (URI), INIST/CNRS, UPS76
[3]Groupe Bioinformatique, Génétique Oncologique, IGR/CNRS, UMR 8125

Jean Royauté
Laboratoire d'Informatique Fondamentale de Marseille (LIF)
UMR 6166 CNRS - Univ. de Provence - Univ. de la Méditerranée
Parc scientifique et technologique de Luminy
CASE 901, 163 Avenue de Luminy
F-13288 Marseille Cedex
Jean.Royaute@lif.univ-mrs.fr

Résumé

Une approche terminologique et infométrique de fouille de texte a été réalisée afin de mettre en évidence des relations entre gènes et protéines impliqués dans les cancers de la thyroïde, à partir d'un corpus de résumés issu de la base de données bibliographiques Medline. L'originalité de cette approche repose sur une indexation contrôlée et structurée en classes sémantiques à partir de vastes ressources hétérogènes provenant de deux bases terminologiques en médecine, biologie et génétique : UMLS et LocusLink. Cette indexation tient compte de la spécificité des termes : nomenclatures biochimiques, acronymes de gènes, aberrations chromosomiques ou encore variantes linguistiques de termes. Deux méthodes de classification complémentaires ont été appliquées sur cet index structuré, permettant de mettre en évidence : (i) trois pathologies de la thyroïde : les cancers médullaires (MTC), papillaires (PTC) et des dysfonctionnements du système immunitaire, (ii) un réseau lexical dense de gènes co-occurrents qui a pu être rapproché de ces pathologies.

Mots Clef

Approche terminologique et infométrique, fouille de texte, linguistique informatique, terminologie, analyse de données et classification.

Abstract

We present a terminological and bibliographic text mining process directed toward the search for relationships between genes and proteins involved in thyroid cancer from a corpus of abstracts extracted from the Medline bibliographic database. The originality of our approach lies in an indexation controlled and structured in semantic classes, from two large heterogeneous lexical resources in the fields of medicine, biology and genomics: UMLS (Unified Medical Language System) and LocusLink. This indexing takes into account the specificity of terms: biochemical nomenclatura, gene acronyms, chromosomic aberrations, or linguistic variants of terms. Two complementary classifications were used in this structural index enabling to highlight: (i) three thyroid pathologies: medullary carcinomas (MTC), papillary carcinomas (PTC) and pathologies which are implicated in immune system dysfunction; (ii) a dense lexical network of coocurrent genes which is linked to these pathologies.

Keywords

Terminological and informetric approach, text mining, computational linguistics, terminology, automatic clustering.

1 Introduction

Nous présentons un processus de fouille de texte orienté vers la recherche de relations entre gènes et protéines dans les cancers de la thyroïde, à partir d'un corpus de résumés de la base de données bibliographiques Medline. Ce processus met en œuvre un ensemble de traitements associant des méthodes de linguistique et d'analyse de données. Les traitements linguistiques reposent sur une indexation à partir de lexiques contrôlés et non sur une extraction terminologique à partir de patrons syntaxiques (Daille et al. 2001). De plus, l'indexation tient compte de la spécificité des termes de ce domaine scientifique :

termes des nomenclatures biochimiques présentant des difficultés particulières d'alignement entre textes et lexiques, gènes sous forme d'acronymes dont il faut s'assurer qu'ils ne sont pas des faux-amis, notations codifiées d'aberrations chromosomiques que l'on cherche à décomposer et traduire et enfin termes considérés comme variantes linguistiques des termes du lexique. Les termes extraits sont ensuite utilisés pour classifier les documents suivant un algorithme de classification non-hiérarchique (NEURODOC) qui réalise une classification thématique des gènes et maladies et suivant un algorithme de classification hiérarchique reposant sur un réseau lexical de termes co-occurrents que sont les noms de gènes (SDOC).

L'originalité de notre approche procède : (i) de l'utilisation de vastes ressources hétérogènes provenant de deux bases lexicales de médecine, biologie et génétique (UMLS et LocusLink), dont la structure interne permet de générer, pour chaque notice bibliographique, non pas une liste plate de termes mais un index contrôlé et structuré où chaque terme appartient à une ou plusieurs catégories sémantiques ; (ii) des méthodes de classification complémentaires exploitant la souplesse de cette indexation pour associer des réseaux de gènes à des pathologies.

Nous décrivons dans la section 2 l'ensemble des traitements. La section 3 analyse les résultats d'indexation ainsi que l'apport des traitements terminologiques dans le processus global. Enfin la section 4 est consacrée aux classifications et à l'interprétation des processus biologiques mis en évidence.

2 Traitement terminologique et infométrique

La chaîne de traitements se décompose en trois grandes étapes représentées dans la figure 1. La première étape est une phase de pré-traitement des données qui concerne l'acquisition du corpus bibliographique par interrogation de la base de données Medline, l'acquisition et les traitements de l'UMLS et LocusLink, ainsi que le formatage du corpus et des ressources en SGML. Précisons que le corpus textuel porte sur la génomique et la protéomique de la glande thyroïde et de ses cancers et regroupe 6 256 références de langue anglaise extraites de la base de données Medline, datées de 1965 à septembre 2001.

La seconde partie de la figure présente les traitements terminologiques mis en œuvre. L'indexation automatique est réalisée par trois modules adaptés aux spécificités de ces corpus: ILC, IRC3, et ANTXT. L'ensemble de ces modules fournit un index contrôlé et structuré où à chaque terme est associé un préférentiel, la séquence textuelle trouvée dans le texte, les éventuelles synonymes et formes variantes, ainsi que la/les classe(s) sémantique(s) attribuée(s) à ce terme à partir des catégories sémantiques de l'UMLS.

La troisième partie indique les traitements de classification. La phase d'analyse de données met à profit deux méthodes de classification complémentaires associées à deux méthodes de cartographie. Cette phase permet de regrouper les documents en classes et d'organiser les termes en réseaux lexicaux au sein des classes. Il existe une forte interactivité entre ces deux dernières étapes, plusieurs validations de vocabulaire et classifications pouvant être nécessaires avant l'analyse réalisée par les experts. Ultime étape du processus, cette analyse consiste en une validation des classes et de leur homogénéité, une verbalisation de leur contenu, et l'interprétation de leurs proximités relatives sur la carte ainsi que des réseaux lexicaux obtenus.

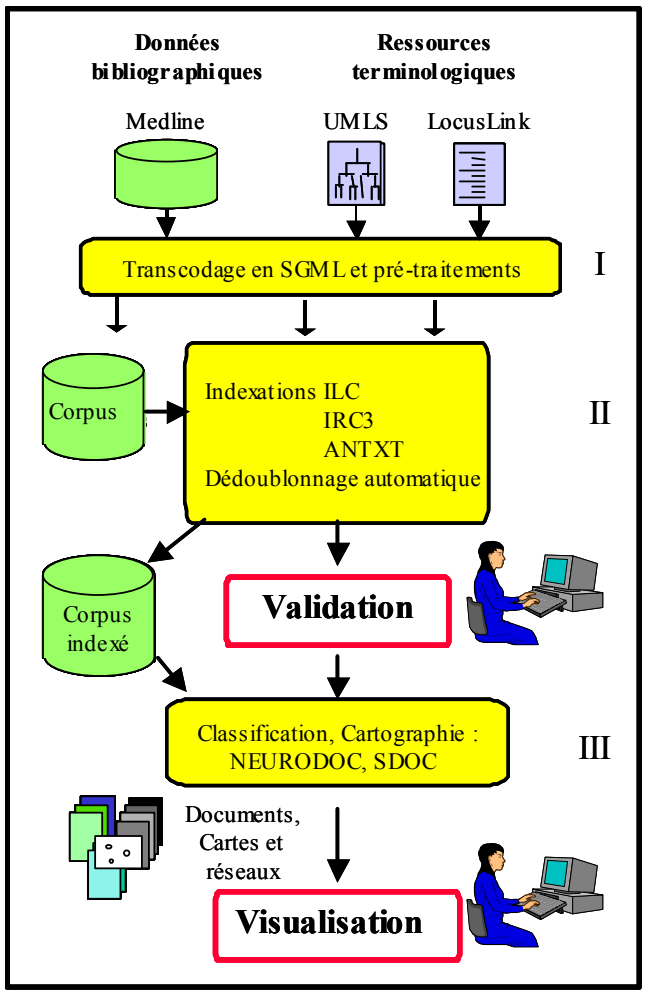

Figure 1: Chaîne des traitements terminologiques et infométriques.

2.1 Les données terminologiques et leur pré-traitement

Le corpus textuel porte sur la génomique et la protéomique de la glande thyroïde et de ses cancers et regroupe 6 256 références de langue anglaise extraites de la base de données Medline, datées de 1965 à septembre 2001.

Une ressource lexicale de 1 005 989 termes a été constituée pour l'indexation à partir de LocusLink et UMLS. LocusLink est une base de données contenant des

informations sur les gènes et les protéines, essentiellement humains : acronymes, noms, synonymes, définitions, structures et fonctions selon la nomenclature mise au point par « Gene Ontology ». Le fichier texte, déchargé en septembre 2001, a permis de rassembler une liste de 37 880 acronymes différents correspondant à 20356 noms de gènes.

UMLS est considéré comme la ressource terminologique la plus importante et la plus complète dans le domaine de la biologie et de la médecine. L'édition utilisée (de janvier 2001) est constituée, pour l'anglais, de 797 359 concepts exprimés sous 1 454 759 formes variantes et organisée en 134 catégories sémantiques dont 75 pertinentes pour notre corpus. Nous avons effectué un regroupement de ces catégories en 9 macro-catégories dont les plus importantes sont : « Organism », « Anatomy », « Biology, Physiology, and Cellular Biology », « Biochemistry and Molecular Biology », « Diseases », « Chemicals and Drugs », etc. Les 797 359 concepts et 1 454 759 termes en anglais ont été fusionnés en un fichier unique de structure SGML. Les différents champs de cette structure reprennent et organisent les informations présentes dans la base. Les caractéristiques structurelles les plus importantes concernent le statut du terme : préférentiel, synonyme, forme variante, synonyme temporaire, et les types de variantes : variante de casse, variante portant sur l'ordre des mots, variante au singulier et variante au pluriel. Ce mode de stockage a rendu possible des traitements automatiques de filtrage afin de rendre la ressource lexicale plus appropriée pour le processus d'indexation.

2.2 Les différents traitements d'indexation

Un processus complexe d'indexation contrôlée à partir de termes simples et complexes a été mis en place afin de tenir compte des particularités du vocabulaire de ce corpus. Nous avons dû envisager trois types différents d'indexation propres à la nature de la terminologie recherchée. C'est ainsi que pour les composés chimiques et biochimiques, qui échappent aux représentations du langage naturel, nous avons développé un outil spécifique de reconnaissance (IRC3, cf. 2.2.1). L'identification des acronymes de gènes, d'une part, et la recherche des aberrations chromosomiques, d'autre part, ont également nécessité un traitement spécial et un outil approprié (ANTXT, cf. 2.2.2). Pour tous les termes comportant une description linguistique dans les thesauri, nous avons utilisé la plate-forme ILC (cf. section 2.2.3). Enfin, la dernière phase d'indexation a été suivie d'une phase de contrôle et de filtrage des termes. Un contrôle de la qualité de l'indexation automatique a permis de supprimer des notices les termes non pertinents selon un critère linguistique ou sémantique.Un filtrage des termes les plus fréquents a éliminé ceux qui risqueraient de perturber la phase ultérieure de classification.

2.2.1 Indexation des composés chimiques et biochimiques

Le programme IRC3 (Indexation par Recherche et Comparaison de Chaînes de Caractères) a été conçu pour les besoins du projet afin d'indexer les termes de certaines catégories sémantiques de l'UMLS mal gérés par ILC, car trop éloignés du langage naturel. Il s'agit d'un programme simple robuste de reconnaissance de composés chimiques et biochimiques. Il part de l'observation que la plupart des échecs de reconnaissance de telles chaînes proviennent de l'absence ou présence d'espaces à plusieurs endroits de la chaîne de caractères dans le texte par rapport au lexique. IRC3 n'utilise donc pas d'heuristiques sophistiquées reposant sur une reconnaissance s'appuyant sur la variabilité syntagmatique et contextuelle de telles expressions (Cohen et al, 2002). Dans une première étape, ce programme écrit en *perl*, crée une table à partir des lexiques. Des espaces sont insérés entre tous les caractères non alphanumériques des chaînes de caractères des termes, qui sont également transformés en minuscule La même opération est réalisée sur les textes. Le programme réalise l'indexation par une recherche dichotomique de tous les mots du texte un par un dans la table ainsi constituée. Les synonymes sont renvoyés vers les préférentiels correspondants à l'aide d'une table de hachage. Ces traitements permettent dans la portion de texte suivant : *... labeling, 4', 6-diamidino-2-phenylindole staining, DNA laddering ...,* de repérer le composé « 4',6-diamidino-2-phenylindole », présent dans le lexique, qui diffère du texte par l'absence d'espace dans deux positions.

2.2.2 Indexation des acronymes de gènes et des aberrations chromosomiques

Le programme ANTXT (ANalyse de TeXTe), écrit en *perl,* présente une double fonctionnalité : repérer les noms de gènes sous forme d'acronymes (en minimisant les risques de confusion avec d'autres entités) ou identifier les expressions relatives à des anomalies cytogénétiques.

translocation	/^t\([XY1-9]/) || /^t\([12][0-9]/
telomeric association	/^tas{0,1}\([XY1-9]/) || /^tas{0,1}\([12][0-9]/
insertion	/^[\+]{0,1}ins\(/
isochromosome	/^[\+]{0,1}i\(/
inversion	/^[\+]{0,1}inv\(/
deletion	/^[\+]{0,1}del\(/
derivative	/^[\+]{0,1}der\([1-9XY]\)/) || /^der\([1-2][0-9]\)/
ring	/^r\([1-9XY]\)/) || /^r\([1-2][0-9]\)/
dicentric	/^[\+]{0,1}[i]*dic\(/
duplication	/^[\+]{0,1}dup\(/
location	/^[1-9XY][pq]/) || /^[12][0-9][pq]/

Tableau 1 : exemple d'expressions régulières permettant d'identifier les entités cytogénétiques

Pour le repérage des acronymes de gène, ANTXT identifie tout mot en majuscules (de 2 lettres minimum) et

le compare à une liste préétablie d'abréviations de noms de maladies, ou à une liste de symboles et d'alias de gènes. Le but ici n'est pas de faire de l'acquisition d'acronymes (Nenadic et al. 2003), (Schwartz et Hearst 2003), à partir des formes développées présentes dans le texte, mais de les différencier des noms de maladies. La liste des abréviations de maladies a été établie à partir de l'expertise de l'ensemble des abréviations (ex : Multiple Endocrine Neoplasia type 2A (MEN2A). La liste des acronymes de gènes a pu ainsi être, en partie, désambiguïsée automatiquement améliorant la précision de l'extraction terminologique.

Par ailleurs, un certain nombre de notations sont utilisées par les biologistes pour décrire de manière concise les anomalies cytogénétiques. La seconde fonctionnalité d'ANTXT consiste à repérer ces anomalies par comparaison avec les patrons ou expressions régulières associés aux différents remaniements chromosomiques.

C'est ainsi que l'expression régulière: /^t\([XY1-9]/) || /^t\([12][0-9]/, appliquée sur la portion de texte : *A novel chromosomal translocation t(3;5)(q12;p15.3) and …*, permet d'indexer une aberration de translocation : t(3;5)(q12;p15.3), portant sur les chromosomes 3 et 5. Le tableau 1 présente quelques-unes des expressions régulières utilisées pour caractériser les anomalies cytogénétiques.

2.2.3 Indexation à partir de traitements linguistiques informatiques

Notre approche terminologique diffère de celle de Nenadic et al. (2003) par le fait que les traitements des variantes de termes reposent exclusivement sur des traitements linguistiques à base de règles et non sur des traitements statistiques. Elle se rapproche des méthodes d'appariement fondées sur une distance entre des mots de sens voisin (Aronson 2001), bien que la micro-syntaxe des termes ne soit pas exploitée. Les traitements linguistiques, réalisés par la plate-forme ILC (Royauté 1999; Royauté et al. 2001), se fondent sur une description syntaxique des termes et leur éventuelle variation en corpus. Ils reposent sur une intégration d'un ensemble de modules linguistiques. Ces traitements, qui exploitent les différents lexiques d'UMLS, permettent d'indexer les documents à partir de termes appartenant à ces ressources (Jacquemin et al. 2002; Daille et al. 2001; Royauté 1999).

Les termes sont identifiés sous des formes identiques à celles de leurs enregistrements ou sous des formes variantes dont le sens est préservé (Jacquemin 1994; Jacquemin et Royauté 1994; Jacquemin et Tzoukermann 1999). Outre les variantes flexionnelles, on distingue quatre types de variations. (i) La variation d'insertion concerne l'ajout de tout mot à l'intérieur du groupe nominal ; par exemple la séquence *immune system function*, associée au terme *Immune function*. (ii) La variation de coordination concerne toutes les formes coordonnées de mots (adjectifs ou noms) à l'intérieur du groupe nominal; par exemple la séquence *skin or subcutaneous tissue* associée au terme *Skin tissue*. (iii) La variation de permutation implique tous les mots ou groupes de mots pouvant permuter autour d'un élément pivot (prépositions ou séquences verbales); par exemple la séquence *protection of thyroid cancer cell* est associée au terme *Cell protection.* (iv) Enfin, la variation morpho-dérivationnelle prend en compte les propriétés linguistiques de mots pouvant être dérivés en d'autres mots de catégories grammaticales différentes (*abdomen /abdominal* pour les dérivations nom/adjectif, *abort/abortion* pour le cas des dérivations verbe/nom, etc.). Ce type de variations peut être associé à des variations syntaxiques comme le montre la séquence *transplantation of monodispersed rat thyroid cells,* transformée du terme *Cell transplant*.

Le processus d'indexation automatisée nécessite que chaque terme soit étiqueté grammaticalement et lemmatisé. Cette phase est réalisée avec le TreeTagger[1] (Schmid 1994). Cela permet de transformer les termes selon deux ensembles de règles en *PATR-II* : les règles sur les mots et sur les termes, dans le formalisme de l'analyseur FASTR (Jacquemin 1994; Jacquemin et Tzoukermann 1999). Les règles sur les mots donnent la catégorie de partie de discours de chaque mot, sa forme lemmatisée et ses liens avec d'autres catégories morpho-dérivationnelles extraites de la base de données CELEX[2]. Les règles sur les termes décrivent la structure syntaxique du terme ainsi que les informations sur les parties de discours de chaque mot du terme. Le corpus doit subir également une transformation pour laquelle chaque mot est étiqueté avec le TreeTagger puis transformé en PATR-II. L'indexation porte donc sur ces deux ensembles de données transformées : lexiques et corpus textuel. Un ensemble de règles particulières, nommées métarègles, permet d'identifier les variantes de chaque terme. Ces métarègles opèrent sur les règles des termes et décrivent à quelle condition la transformation d'un terme en sa variante est possible pour son indexation.

2.3 Classification et cartographie

L'étape de classification et cartographie permet de structurer une grande masse d'information dans le but de mettre en évidence les relations entre les gènes et les pathologies tumorales de la thyroïde. Deux méthodes complémentaires sont utilisées : une classification thématique à partir des noms de gènes et maladies avec l'application NEURODOC et une classification reposant sur un réseau lexical de noms de gènes co-occurrents (application SDOC) afin de préciser quelles peuvent être les relations entre gènes impliqués dans les maladies.

---

[1] Le TreeTagger à été développé par Helmut Schmid à l'Université de Stuttgart. (http://www.ims.uni-stuttgart.de/~schmid/)

[2] CELEX est une base de données lexicales conçue par le « Centre for Lexical Information, Max Plank Institute for Psycholinguistics, Nijmegen.» (http://www.kun.nl/celex/).

2.3.1. NEURODOC

NEURODOC applique l'algorithme des “k-means axiales” comme méthode de classification associée à une Analyse en Composantes Principales (ACP) pour la représentation des classes obtenues dans un espace bidimensionnel (Lelu et François 1992). Les “k-means axiales” (Lelu 1993), variante de l’algorithme des K-means (MacQueen 1967), est une méthode de classification non hiérarchique qui construit des classes recouvrantes dont les éléments, documents et mots, peuvent appartenir à plusieurs classes à la fois et sont ordonnés selon un degré de ressemblance au type idéal de chaque classe.

L'ensemble des classes générées est projeté sur un plan par un ACP constituant ainsi une carte des thèmes qui permet d’apprécier les positions relatives des classes.

Une Analyse en Composantes Connexes (ACC), fondée sur la théorie des graphes, qui calcule les connexions entre les classes 2 à 2, complète l’analyse des “k-means axiales” en dessinant les liaisons inter-classes avec un trait d’intensité variable, selonl’intervalle associé (Polanco et François 2000).

2.3.2. SDOC

SDOC (Grivel et al. 1995) est une réalisation informatique de la "*méthode des mots associés*" (Callon et al. 1986). Cet algorithme fondé sur la cooccurrence des mots-clés met en évidence la structure de leurs relations (réseaux lexicaux). La notion de cooccurrence permet de définir une proximité entre mots-clés : deux termes figurant ensemble dans les documents étant considérés comme proches. L’emploi d’un indice statistique permet de normaliser la mesure de l’association entre deux mots clés.

Nous utilisons ici l’indice dit *d’Equivalence* E_{ij} (Callon et al. 1986) dont les valeurs varient entre 0 et 1:

$$E_{ij} = \frac{C_{ij}^2}{(C_i * C_j)}$$

où C_{ij} est le nombre de cooccurrences des mots-clés *i* et *j*, C_i la fréquence du mot-clé *i* et C_j la fréquence du mot-clé *j*. Une fois le réseau complet calculé, ***SDOC*** applique un algorithme de *Classification Ascendante Hiérarchique* (CAH) dit *du simple lien* (*single link clustering*), afin de construire des classes ou clusters de mots, proches les uns des autres, n'excédant pas une taille maximale. Un cluster est donc constitué de mots associés les uns aux autres (*associations internes*), les clusters pouvant avoir des relations entre eux (*associations externes*). Après le processus de classification des mots-clés, les documents sont affectés aux clusters.

Les résultats des classifications NEURODOC et SDOC sont mis à disposition des utilisateurs sous l’interface de travail VISA pour la phase de validation. Les cartes ne sont pas seulement un moyen de visualisation, elles représentent aussi une méthode d'analyse dans la mesure où elles permettent d'évaluer la position des thèmes entre eux dans un espace géométrique de représentation.

3 Analyse des résultats d’indexation

L’analyse des résultats montre la complémentarité des différentes méthodes utilisées pour l’indexation des notices bibliographiques. Parmi les 6 059 mots clés indexant les documents, près d’un tiers n’indexe qu’un seul document, soit 2 723 termes. La proportion de mots clés de fréquence 1 est de près de 50% dans les catégories « Diseases » (C06) et « Protein » (C05), et de plus de 80% pour les aberrations chromosomiques (122 sur 148). Les termes de biochimie sont les plus nombreux (1 554) et indexent 61% des documents. Les noms de maladies, moins nombreux (955), indexent le plus de documents. La moyenne de mots clés par document est de 13.

Le calcul de la précision des différentes indexations a été effectué à partir des opérations de filtrage des termes bruités. Ainsi, sur 37 619 séquences textuelles pouvant être reliées à un terme d’indexation, 25 846 ont été conservées, ce qui donne une précision globale de 0,69. Cette précision est fonction des ressources et de l’outil d’indexation utilisés. Elle est de 1 pour les aberrations chromosomiques, de 0,47 pour la recherche d’acronymes de gènes avec le fichier LocusLink, de 0,73 avec ILC et de 0,91 avec IRC3. On remarque aussi que la richesse des relations de synonymie présentes dans l’UMLS a permis de ramener 38% des séquences textuelles obtenues par ILC et IRC3 réunis (11 545 sur les 30 652). Parmi celles-ci, 7 963 ont été conservées par validation humaine, soit une précision de 0,69.

3.1 L’indexation avec ILC

Nous avons vu que la précision de ILC était de 0,73, ce qui correspond à 26 108 séquences textuelles reconnues pour 18 976 conservées. Ramené au nombre de termes différents, ce chiffre correspond à une précision de 0,72, soit 10 108 préférentiels extraits pour 7 304 conservés. L’apport des traitements linguistiques réalisés par ILC est important puisque plus de la moitié des séquences textuelles ont été obtenues par une variation terminologique (13 550 soit 52% du total). L’évaluation de l’indexation se mesure aussi par les regroupements de termes opérés avec les liens de synonymie propres à UMLS et qui viennent s’ajouter aux regroupements linguistiques des variations terminologiques. C’est ainsi que 38% des termes identifiés sont des synonymes. Cependant la synonymie peut poser problème et les liens ne sont pas toujours justifiés quand on les projette sur des textes. C’est ainsi que la précision, de 0,77, pour des termes non reliés à un synonyme passe à 0,66 sous l’effet conjugué de mauvais liens de synonymie et de variations inadéquates.

Si on analyse ces résultats par rapport aux types de variations traités, on observe une forte proportion de variantes d’insertion (37%), viennent ensuite les variantes de permutation (24%), les variantes

morpho-dérivationnelles (21%), puis les variantes de coordination (12%), la précision variant de 0,56 à 0,87. On remarque le très bon score de la permutation (0,87) qui dépasse la coordination (0,85), réputée plus filtrante que les autres variantes, les variations morpho-dérivationnelles, présentant le taux de précision le plus faible (0,56). Il y a deux explications à ce faible taux. D'une part, il s'agit des variantes les plus complexes, cumulant deux difficultés linguistiques : la reconnaissance de formes syntaxiques équivalentes et l'identification correcte et non artefactuelle d'un lien morphologique entre deux catégories syntaxiques différentes. D'autre part, il a été observé des appariements morphologiques qui posent problème avec la base CELEX, notamment avec les formes préfixées, qui tendent à modifier le sens de façon incontrôlée. Augmenter la précision implique un filtrage plus strict de la base lexicale. Par ailleurs, ces chiffres de précision varient en fonction des catégories sémantiques. Les termes relevant de concepts généraux, peu spécialisés dans le domaine de l'étude, tels que « Biology, physiology and cellular biology » ou « Cells » varient entre 0,10 et 0,50. En revanche les termes provenant de catégories très spécialisées, tels que ceux de « Anatomy » ou « Diseases » ont une meilleure précision.

3.2 L'indexation avec IRC3

Les résultats d'indexation d'IRC3 montrent le rôle important des termes de composés biochimiques. Ils représentent 4 544 séquences textuelles et apportent 3 188 préférentiels supplémentaires à l'indexation des documents, soit 29% du total de l'indexation**.** La précision, avec un taux de 0,91 est très bonne. Les formes éliminées sont souvent dues à un mauvais renvoi de synonymie à partir d'un acronyme. Sur 3 567 préférentiels, 3 188 sont conservés, soit une précision de 0,89. L'apport de la synonymie est moins riche qu'avec les autres catégories. Seulement 27% des termes préférentiels sont obtenu par synonymie avec une précision de 0,89. Ces noms de composés chimiques et biochimiques pour la plupart, sont gérés par des nomenclatures sur lesquelles s'appuie l'UMLS pour constituer son vocabulaire, et auxquelles se réfèrent également les auteurs. Ils sont peu soumis à variations, et le préférentiel UMLS correspond souvent au terme officiel utilisé par les auteurs.

3.3 L'indexation des gènes, protéines et des anomalies cytogénétiques

Nous analysons ici l'indexation des gènes ou protéines avec les ressources : LocusLink et UMLS, et les outils : ILC, IRC3 et ANTXT. Nous comptabilisons ainsi 939 protéines dans 3 826 notices, dont 454 (48%) de fréquence 1. La contribution des deux ressources est équitablement répartie puisque 477 proviennent de LocusLink et 462 d'UMLS. A partir de LocusLink et par recherche des acronymes de gènes (programme ANTXT), la précision est, ainsi que nous l'avons vu, de 0,43, le nombre d'occurrences identifiées étant de 6 759. Ce nombre correspond à 477 acronymes préférentiels de différentes protéines. Avec UMLS, nous obtenons 10 693 occurrences pouvant correspondre également à des protéines. 462 protéines différentes ont été conservées. Enfin, 140 protéines indexées par l'UMLS sont renvoyées vers un acronyme LocusLink. La précision de l'indexation des aberrations chromosomiques, avec un taux de 1 est excellente. La méthode a permis de retrouver dans 68 notices 204 occurrences d'aberrations correspondant à 148 types différents.

4 Classification et interprétation des processus biologiques

Afin de caractériser les relations entre les gènes (et/ou protéines) et les pathologies tumorales de la thyroïde, nous avons réalisé les classifications automatiques spécialisées au vocabulaire d'indexation associé aux catégories sémantiques : *« genes and proteins »* et *« diseases »*. Notre approche diffère d'autres études (Pillet 2001) où des phrases dénotant des interactions entre gènes sont recherchées à partir d'un indice exploitant leur cooccurrence et la présence de certains « mots déclencheurs ». D'autres méthodes utilisent des transducteurs (Poibeau 2001) ou des méthodes d'apprentissage à partir de patrons prédicatifs (Bessières et al. 2001) pour mettre en évidence de telles relations. Cependant ces méthodes ne s'intéressent qu'à un seul type d'interaction : gène/gène ou gène/protéine et ne semblent pas les mieux adaptées pour relier des gènes aux pathologies. Les réseaux lexicaux que nous produisons ressemblent dans leur forme à ceux de SUISEKI (Blaschke et Valencia 2001) pour la recherche d'interaction entre gènes. Ils sont construits uniquement à partir de la cooccurrence de noms de gènes, révélée par nos traitements linguistiques d'indexation.

Nous avons réalisé une première analyse en sélectionnant automatiquement dans l'indexation les noms de gènes, protéines et pathologies et en appliquant l'algorithme NEURODOC. Nous avons ensuite créé le réseau d'association des noms de gènes et protéines obtenu par l'application de l'algorithme SDOC sur l'index *« genes and proteins ».* Puis, ce réseau a été interprété en utilisant les annotations associées aux classes disponibles sur l'interface utilisateur (listes de maladies associées aux classes, liens vers les bases factuelles et ontologies).

Pour illustrer les résultats obtenus, nous présentons ci-après une analyse de la carte obtenue avec la classification NEURODOC associant gènes, protéines et pathologies ; nous nous focalisons ensuite sur le réseau lexical de la classe « RET » obtenue avec la classification SDOC.

4.1 Classification obtenue à partir de l'indexation des gènes, protéines et maladies

Cette classification regroupant deux catégories sémantiques de mots clés a été réalisée pour révéler les principales associations entre gènes et maladies. L'analyse de la composition respective de chaque classe permet de différencier trois ensembles sur la carte de la figure 2. Trois groupes bien différenciés par l'analyse en composantes principales (ACP) se positionnent sur des parties différentes de la carte, l'analyse en composantes connexes (ACC) révélant les plus fortes connexions à l'intérieur de chaque groupe.

Groupe 1 : Ce groupe est formé de 8 classes, dont 4 sont plus fortement liées par l'ACC (*RET, Pentagastrin, Multiple Endocrine* et *Hyperplasia*). Les classes concernent des pathologies liées à des mutations ponctuelles du gène RET dont la principale est le cancer médullaire de la thyroïde (MTC). Elles sont plutôt orientées cliniques avec des marqueurs de diagnostic tel que la pentagastrine, la thyrotropine et la calcitonine.

Groupe 2 : Ce groupe est formé de 7 classes dont 4 sont plus fortement liées par l'ACC (*RAS, TP53, Carcinogenesis* et *Tumor progression*). Ces classes concernent le deuxième type de pathologie tumorale en relation avec des translocations et des fusions chromosomiques concernant le gène RET : les cancers papillaires (PTC) différenciés ou pas. L'ensemble est plutôt orienté « pathogénie moléculaire » et regroupe les oncogènes. Une connexion entre les groupes 1 et 2 (*RET* et *carcinogenesis)* s'explique par le rôle que joue le gène *RET* dans ces deux types de cancers de la thyroïde.

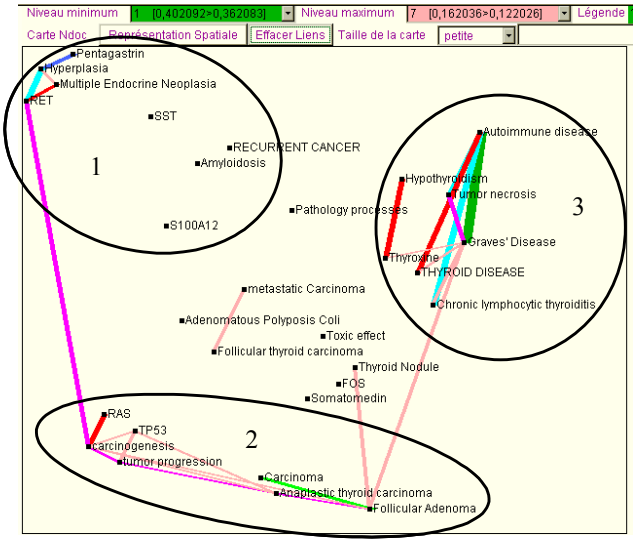

Figure 2: Carte globale obtenue par classification Neurodoc réalisée à partir de l'index « Gènes, Protéines et Maladies »

Groupe 3 : Ce groupe est formé de 7 classes dont 4 sont plus fortement liées par l'ACC (*Autoimmune Disease, tumor Necrosis, Grave's disease* et *Chronic lymphocytic thyroiditis)*. Il associe un groupe de pathologies de la thyroïde en relation avec des dysfonctionnements du système immunitaire.

4.2 Classification réalisée à partir de l'indexation des gènes et protéines

La classification décrite ci-dessus a mis en évidence l'implication du gène RET dans les deux pathologies cancéreuses de la thyroïde : les PTC et MTC. Nous allons donc étudier dans cette section le sous-réseau associé à ce gène, il se répartie entre deux classes : « RET » et « VIM », ces deux classes étant reliées par une association externe : « RET » - « Calcitonine », soulignée d'un trait plus épais sur les figures 3 et 4.

Outre, ce lien externe, la classe « RET » est formée de 2 sous-ensembles distincts présentés dans la figure 3. Ils montrent deux types de processus biologiques dans lesquels est impliqué ce gène : les mécanismes moléculaires des cancers papillaires (sous-réseau A), et la régulation neuroendocrine par des voies de signalisation intercellulaire (facteurs de croissance neuronaux et leurs récepteurs) (sous-réseau B).

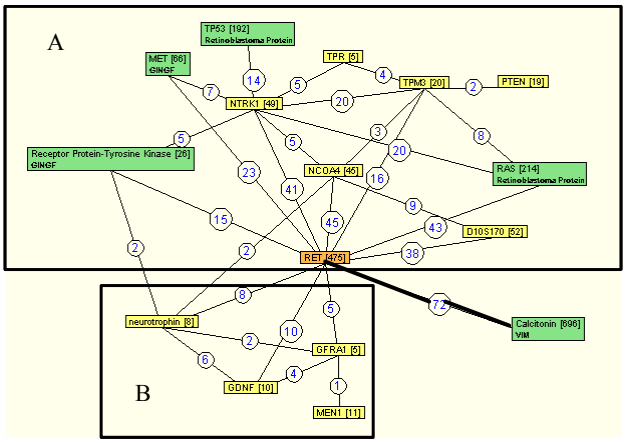

Figure 3 : Réseau des noms de gènes et protéines de classe « RET » obtenue avec la classification SDOC

Sous-réseau A : Avec NCOA4, NTRK1, PTEN, D10S170, TPM3 et TPR, une partie du réseau de relations connues de RET avec d'autres gènes dans le cadre des PTC est ici reconstituée. Ce sous-réseau se prolonge par les liens vers les classes *GINGF* et *Retinoblastoma protein*. Cette portion de réseau porte sur les mécanismes moléculaires des PTC (réarrangements géniques, fusion de gènes), et nous renseigne sur les principales interactions entre protéines ou gènes décrites dans la cadre de l'oncogénèse des PTC (réarrangement NTRK1 et TPM3 donnant l'oncogène TRK, RAS, TP53 et MET). Les gènes de ce sous-réseau se retrouvent dans les classes situées sur la partie gauche du groupe 2 de la carte Neurodoc portant sur cette maladie (RAS, TP53,Carcinogenesis).

Sous-réseau B : Ce micro-réseau relie le gène RET aux molécules GDNF, GFRA1 (ligands de RET). La présence de la neurotrophine confirme l'implication de RET dans la

régulation neuroendocrine par des voies de signalisation intercellulaire.
Le réseau de la classe « VIM » est présenté sur la figure 4, il est divisé en 2 sous-réseaux chacun étant impliqué dans une pathologie tumorale de la thyroïde.
Sous-réseau C : Ce sous-réseau situé sur la partie supérieure de la figure intègre la relation « RET » - « Calcitonin » et associe ce gène à Pentagastrin, SST, Amyloid qui se retrouvent dans les classes du groupe 1 de la carte Neurodoc (Pentagastrin, Hyperplasia, SST, recurrent cancer, amyloidosis, S100A12). C'est un réseau de gènes en relation avec les cancers médullaires (MTC).
Sous-réseau D : Ce sous-réseau autour du gène VIM concerne également les cancers papillaires (PTC). Les gènes associés à ce sous-réseau se retrouvent dans la description des classes situées sur la partie droite de la carte avec entre autres les classes « Anaplastic thyroid carcinoma » et « Follicular Adenoma » du groupe 2.

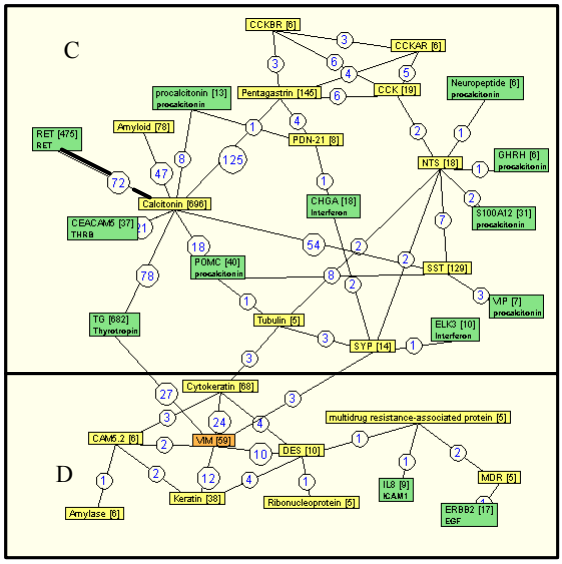

Figure 4 : Réseau des noms de gènes et protéines de classe « VIM » obtenue avec la classification SDOC

Ces deux classifications se complètent donc pour décrire les réseaux de gènes et protéines associés au gène « RET » dans les différentes pathologies cancéreuses de la thyroïde.

5 Conclusion

Dans ce projet nous avons cherché à mettre en évidence les relations entre gènes (ou protéines) et les pathologies tumorales de la thyroïde, par des méthodes terminologiques de TALN et des méthodes d'analyse de données. L'originalité de notre approche découle de l'utilisation, pour l'indexation, de vastes ressources hétérogènes et structurées provenant de deux ressources lexicales de médecine, biologie et génétique. Cela a permis d'obtenir pour chaque notice bibliographique un index contrôlé et structuré où chaque terme appartient à une ou plusieurs classes sémantiques permettant de faire des classifications très ciblées. Nous nous sommes donc intéressés plus particulièrement à deux catégories sémantiques issues de notre indexation : les maladies et les gènes, pour produire des classifications. Chaque classe est annotée par les différentes maladies et les gènes cités dans les articles qui les constituent. La classification réalisée avec NEURODOC nous a permis de mettre en évidence des groupes qui se structurent autour de trois types importants de pathologies de la thyroïde : les cancers médullaires (MTC), les cancers papillaires (PTC), et enfin les pathologies liées à des dysfonctionnements du système immunitaire. Ces classes ont confirmé l'importance du gène RET. Nous avons donc analysé la partie du réseau lexical obtenu avec SDOC formé autour de l'intitulé de ce gène. La structure interne de ce réseau a également justifié le rôle de ce gène dans les deux principaux cancers de la thyroïde tout en précisant les relations qu'il établit avec les autres gènes.
Les informations et relations que nous mettons en évidence, connues des experts, n'apportent pas, à proprement parler, d'éléments nouveaux et/ou des rapprochements inattendus. Notre ambition, dans l'état actuel du projet, a été la mise en place d'une méthodologie d'analyse de données afin d'en faciliter l'analyse experte. Etre capable, par nos méthodes, de révéler une information manipulable et interprétable par les experts à partir de connaissances qui leur sont familières est pour nous l'étape préliminaire à des travaux plus ambitieux de « découverte de connaissance ».

Remerciements

Ce travail a été financé par le CNRS, l'INRA, l'INRIA, et l'INSERM dans le cadre du projet inter-EPST « Bionformatique ». Les auteurs remercient les cliniciens et biologistes de l'Institut Gustave Roussy, experts avec lesquels ils ont eu de fructueux échanges tout au long de ce travail : Jean-Michel Bidart, Bernard Caillou, Martin Schlumberger.

Bibliographie

AR. Aronson, *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program*, Proc AMIA Symp. 2001;:17-21.
P. Bessières, A. Nazarenko, C. Nedellec., *Apport de l'apprentissage à l'extraction d'information: le problème de l'identification d'interactions géniques*, In Conférence Internationale sur le Document Electronique, (CIDE'2001), Toulouse, octobre 2001.
C. Blaschke et A. Valencia., *The Potential Use of SUISEKI as a Protein Interaction Discovery Tool*, In Genome Informatics, 12 123-134, 2001.
M. Callon, J. Law, A. Rip., (editors) *Mapping the Dynamics of Science and Technology*. London, MacMillan Press, 1986.
K. B. Cohen, A. E. Dolbey, G. K. Acquaah-Mensah and L. Hunter., (2002). *Contrast and variability in gene*

names. Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, 14-20, 2002.

B. Daille, J. Royauté, X. Polanco., *Evaluation d'une plate-forme d'indexation de termes complexes*, T.A.L. journal (Le traitement automatique des langues http://www.atala.org/tal/tal.), Special issue in « Information retrieval – oriented natural language processing », Vol. 41, Num. 2, January 2001.

Grivel, P. Mutschke, X. Polanco., "Thematic mapping on bibliographic databases by cluster analysis : a description of the SDOC environment with SOLIS", *Journal of Knowledge Organization*, Vol. 22, 1995, n° 2, p. 70-77.

C. Jacquemin et J. Royauté., *Retrieving Terms and their Variants in a Lexicalised Unification-Based Framework*, Proceedings, 17th Annual International ACM SIGIR., Dublin, 1994.

C. Jacquemin et E. Tzoukermann., *NLP for term variant extraction: Synergy of morphology, lexicon, andsyntax*, In T. Strzalkowski (Ed.), Natural language information retrieval, 1999, p 25-74, Boston, MA: Kluwer.

C. Jacquemin, B. Daille, J. Royauté et X. Polanco., *In Vitro Evaluation of a Program for Machine-Aided ,* Information Processind & Management. Volume 38, Issue 6, Pages 765-792, 2002.

A. Lelu., *Modèles neuronaux pour l'analyse de données documentaires et textuelles*. Thèse de doctorat de l'Université Paris 6., 1993.

A. Lelu et C. François., *Information retrieval based on a neural unsupervised extraction of thematic fussy clusters*, Neuro-Nîmes 92 : Les réseaux neuro-mimétiques et leurs applications, 2-6 novembre 1992, Nîmes, France.

LocusLink : http://www.ncbi.nlm.nih.gov/LocusLink/.

J. Mac Queen., *Some Methods for Classification and Analysis of Multivariate Observation*. Proceedings of the 5th Berkeley Symposium Mathematics, Statsitics, and Probability (1967), pp. 281-297.

G. Nenadic, I., Spasic, S. Ananiadou, *Terminology-driven mining of biomedical literature*, Bioinformatics. 2003 May 22;19(8):938-43.

V. Pillet., *Méthodologie d'extraction automatique d'information à partir de la littérature scientifique en vue d'alimenter un nouveau système d'information*, Thèse de doctorat de l'Université de droit, d'économie et des sciences d'Aix-Marseille, 2000.

Programme inter-EPST « Bionformatique » 2000-2002 : http://biomserv.univ-lyon1.fr/BioInfo/index.php

T. Poibeau., *Extraction d'information dans les bases de données textuelles en génomique au moyen de transducteurs à nombre fini d'états*, in Actes de la Conférence Française de Traitement Automatique de la Langue, (TALN'2001), 2001.

X. Polanco et C. François., *Data Clustering and Cluster Mapping or Visualization in Text Processing and Mining*, in: Dynamism and Stability in Knowledge Organization: proceceedings of the Sixth international ISKO conférence, 10-13 July 2000, Toronto, Canada. Edited by Clare Beghtol, Lynne C. Howarth, Nancy J. Williamson, Ergon Verlag, p. 359-365.

J. Royauté., *Les groupes nominaux complexes et leurs propriétés : application à l'analyse de l'information*, Thèse de doctorat en informatique, Université Henri Poincaré - Nancy I, 19 juillet 1999, 228 pages.

J. Royauté, C. François et D. Besagni., 2001- *Apport d'une méthodologie de recherche de termes en corpus dans processus de KDD : application de veille en biologie moléculaire,* VSST'2001 (Veille Stratégique Scientifique & Technologique), 15-19 Octobre 2001, BARCELONE, ESPAGNE, Organisation FPC/UPC – SFBA –IRIT, Actes I : full paper, p. 49-62.

M. Schlumberger and F. Pacini., *Thyroid tumors,* Nucléo N Editors, 1999.

A.S. Schwartz et M.A. Hearst, *A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text.* Pacific Symposium on Biocomputing 8:451-462, 2003

H. Schmid., *Probabilistic Part-of-Speech Tagging Using Decision Trees*. Proceedings of International Conference on New Methods in Language Processing. September 1994. http://www.ims.uni-stuttgart.de/~schmid/

UMLS Knowledge Sources – Unified Medical Language System US Department of Health and Human Services. National Institutes of Health. National Library of Medicine.– 12th Edition (Janvier 2001).

Voir aussi :

http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML