Relations entre gènes impliqués dans les cancers de la thyroïde[1]

Jean Royauté*, Claire François**, Alain Zasadzinski**, Dominique Besagni**
Philippe Dessen***, Sylvaine Le Minor***, Marie-Thérèse Maunoury***

* Laboratoire d'Informatique Fondamentale (LIF) UMR 6166 CNRS, CASE 901, 163 Avenue de Luminy, F-13288 Marseille Cedex 9
jean.royaute@lif.univ-mrs.fr
** Unité de Recherche et Innovation (URI), INIST/CNRS; UPS76, 2 Allée du Parc de Brabois, 54514 Vandœuvre Cedex
{francois,zasadzin,besagni}@inist.fr
*** Groupe Bioinformatique, Génétique Oncologique, IGR/CNRS; UMR 8125, Institut Gustave Roussy, 39, rue C Desmoulins, 94805 VILLEJUIF Cedex
{dessen,leminor,mtm}@igr.fr

Résumé. Des relations entre gènes et protéines impliqués dans les cancers de la thyroïde ont été mises en évidence par l'analyse d'un important corpus de résumés de la base de données bibliographique Medline. Une approche pluridisciplinaire (biologistes, cliniciens, linguistes et chercheurs en sciences de l'information) a permis l'indexation automatique et l'analyse de ce corpus. L'indexation contrôlée, structurée en classes sémantiques, à partir de vastes ressources hétérogènes (les bases biomédicales et génétiques UMLS et LocusLink), prend en compte la spécificité des termes : nomenclatures biochimiques, acronymes de gènes, aberrations chromosomiques ou encore variantes linguistiques de termes. Les deux méthodes de classification complémentaires appliquées révèlent un réseau lexical dense de gènes cooccurrents autour des trois principales pathologies de la thyroïde : les cancers médullaires, papillaires et des dysfonctionnements du système immunitaire. Les développements apportés aux outils de visualisation interactifs du serveur VISA de l'INIST facilitent lecture et navigation au sein des documents.

1. Introduction

Le processus de fouille de texte réalisé à partir d'un corpus extrait de Medline porte sur la recherche de relations entre gènes et protéines impliqués dans les cancers de la thyroïde. Sa validation a été facilitée par les modes de visualisation et de navigation des nouveaux développements apportés à l'interface. L'originalité de cette approche procède 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

[1] Une autre version de ce travail, intitulée « Approche terminologique et infométrique de fouille de données textuelles dans un corpus sur la génétique des cancers de la thyroïde », sera présentée au colloque RFIA en janvier 2004.

et structuré où chaque terme appartient à une ou plusieurs catégories sémantiques ; des méthodes de classification complémentaires exploitant la souplesse de cette indexation pour associer des réseaux de gènes à des pathologies. La section 2 présente une vue d'ensemble des traitements. La section 3 examine le processus d'indexation, ses résultats et l'apport des traitements terminologiques dans le processus global. Enfin la section 4 est consacrée aux classifications et à l'interprétation des processus biologiques.

2. Processus de fouille

Le processus de fouille se décompose en trois parties (figure 1). La première concerne le prétraitement des données : l'acquisition du corpus bibliographique par interrogation de la base de données Medline, l'acquisition et les traitements d'UMLS et LocusLink, ainsi que le formatage du corpus et des ressources en SGML.

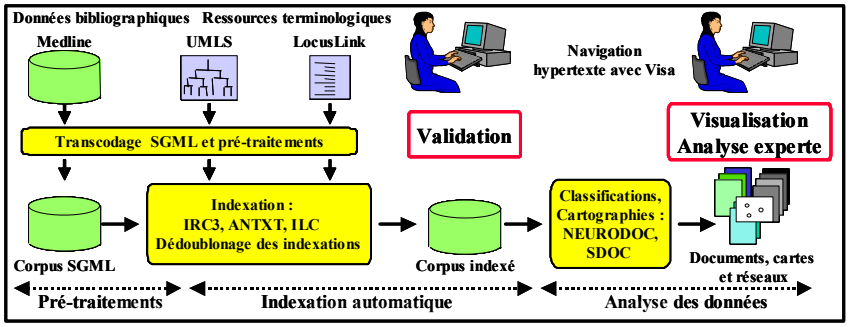


FIG. 1 – *Description fonctionnelle des différents traitements.*

La seconde partie décrit les traitements terminologiques réalisés par trois modules adaptés aux spécificités de ces corpus : IRC3, ANTXT et ILC. Ces modules fournissent un index contrôlé et structuré où à chaque terme est associé un préférentiel, la séquence textuelle trouvée dans le texte, les éventuels synonymes et formes variantes, ainsi que la/les classe(s) sémantique(s) attribuée(s) à ce terme. La troisième partie correspond à la phase d'analyse de données réalisée à partir de deux méthodes de classification complémentaires afin de regrouper les documents en classes et organiser les termes en réseaux lexicaux au sein des classes. Une forte interactivité existe entre ces deux dernières étapes, plusieurs validations de vocabulaire et classifications pouvant être nécessaires avant de présenter les résultats aux experts. L'ultime étape du processus 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.

3. Collecte des données et indexation

Un processus complexe d'indexation contrôlée, dont la précision globale est 0.69 (37 619 occurrences sur 25 846 conservées) a été mis en place pour tenir compte des particularités du vocabulaire de ce corpus. Ce processus a nécessité d'abord de rassembler le corpus et les données terminologiques nécessaires à l'indexation. Ensuite, quatre types différents d'indexation propres à la nature de la terminologie étudiée ont été nécessaires : recherche des

composés chimiques et biochimiques, identification des acronymes de gènes, recherche des aberrations chromosomiques, et enfin repérage des termes ayant une structure linguistique.

3.1. Prétraitement des données

Le corpus porte sur la génomique, la protéomique de la thyroïde et ses cancers. Il regroupe 6 256 références Medline en anglais, datées de 1965 à 200, et pour des besoins de traitement, a été formaté en SGML. Un réservoir de 1 005 989 termes a été constitué pour l'indexation à partir de LocusLink et UMLS. 37 880 acronymes différents correspondant à 20 356 noms de gènes, proviennent de LocusLink, base de données sur les gènes et les protéines. UMLS, la ressource terminologique la plus importante en biologie et médecine, nous a permis de retenir 75 catégories sémantiques pertinentes sur 134, regroupées par la suite en 9 macro-catégories. Les fichiers UMLS ont été fusionnés en un fichier unique au format SGML, reprenant les caractéristiques structurelles les plus importantes telles que 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 permis des traitements automatiques de filtrage afin de rendre la ressource lexicale plus appropriée pour le processus d'indexation.

3.2. Indexation des composés chimiques et biochimiques

IRC3 (Indexation par Recherche et Comparaison de Chaînes de Caractères) est un programme simple et robuste de reconnaissance de composés chimiques et biochimiques. En effet, ces termes incluent des signes de ponctuation : tirets, virgules, parenthèses ou crochets, considérés comme séparateurs par les outils de TALN. IRC3 n'utilise pas d'heuristiques sophistiquées s'appuyant sur la variabilité syntagmatique et contextuelle de ces expressions [Cohen et al., 2002] mais gère l'ajout ou l'omission d'espaces autour des ponctuations, source la plus importante de variation. Dans une première étape, ce programme crée une table à partir des lexiques en insérant des espaces entre tous les caractères non alphanumériques des chaînes de caractères des termes, transformés en minuscule. La même opération est réalisée sur les textes. Le programme réalise l'indexation par une recherche dichotomique de chacun des mots du texte dans la table ainsi constituée.

IRC3 : analyse des résultats d'indexation

Avec 3 188 préférentiels, les composés biochimiques jouent un rôle important dans la description des textes. La précision (0,91) est très bonne (4 544 occurrences ramenées sur 4 131 retenues). 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 plus faible de 0,89.

3.3. Indexation des gènes, protéines et des anomalies cytogénétiques

Le programme ANTXT (ANalyse de TeXTe) 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. Pour le repérage des acronymes de gènes, 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 d'après 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.

ANTXT : analyse des résultats

L'indexation des gènes ou protéines avec les ressources LocusLink et UMLS, et les outils ILC, IRC3 et ANTXT permet de comptabiliser 939 protéines, dont 454 (48%) de fréquence 1. La contribution des deux ressources est équitablement répartie, 477 provenant de LocusLink et 462 d'UMLS. A partir de LocusLink et par recherche des acronymes de gènes (programme ANTXT), la précision est de 0,37, le nombre d'occurrences identifiées étant de 6 759, sur 2 526 retenues. Ce nombre correspond à 477 acronymes préférentiels de protéines différentes. 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. Remarquons, que 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.

3.4. Indexation à partir de traitements linguistiques informatiques

Les traitements linguistiques mis en œuvre reposent sur une indexation à partir de lexiques contrôlés et non sur une extraction terminologique à partir de patrons syntaxiques [Daille et al., 2001]. Notre approche terminologique diffère de celle de Nenadic et al. (2003] par le fait que nos 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 chez cet auteur, 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 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]. Quatre types de variantes sont traités. (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*, *abort/abortion, etc*.).
Le processus d'indexation automatisée repose sur une intégration d'un ensemble de modules linguistiques qui nécessite que chaque terme soit étiqueté grammaticalement et lemmatisé. Cette phase, réalisée avec le TreeTagger[2] [Schmid, 1994], permet de transformer les termes en règles en *PATR-II* selon le formalisme de l'analyseur FASTR [Jacquemin, 1994; Jacquemin et Tzoukermann, 1999], pour lesquelles les informations morpho-dérivationnelles sont extraites de la base de données CELEX[3]. Le corpus doit subir également une transformation similaire où 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 décrivent à quelle condition la transformation d'un terme en sa variante est possible pour son indexation. 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].

ILC : présentation des résultats

La précision globale d'ILC est de 0,73, ce qui correspond à 26 108 occurrences reconnues pour 18 976 conservées, soit 7304 termes préférentiels pertinents. L'apport des traitements 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 opérés avec les liens de synonymie d'UMLS, s'ajoutant aux regroupements des variations terminologiques. Cependant la synonymie peut poser problème et les liens ne sont pas toujours motivés quand on les projette sur des textes. A partir des 38% des termes identifiés comme synonymes, on observe 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%), suivie des variantes de permutation (24%), des variantes morpho-dérivationnelles (21%), puis des variantes de coordination (12%). Leur précision varie de 0,56 à 0,87. En effet, 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 présentent, elles, le taux de précision le plus faible (0,56), ce qui peut s'expliquer de 2 façons. 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

[2] Le TreeTagger à été développé par Helmut Schmid à l'Université de Stuttgart. (http://www.ims.uni-stuttgart.de/~schmid/)
[3] 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/).

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, tous 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.

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 spécifiques aux catégories sémantiques de l'indexation : *« 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 en appliquant le programme Neurodoc. Nous avons ensuite étudié le réseau d'association des noms de gènes et protéines obtenu par application du programme Sdoc. 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 associé au gène « RET » obtenu avec la classification Sdoc.

4.1. Neurodoc : classification et cartographie

Neurodoc applique l'algorithme des "k-means axiales" comme méthode de classification associé 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. Une Analyse en Composantes Connexes (ACC), fondée sur la théorie des graphes, calcule les connexions entre les classes 2 à 2 et complète l'analyse des "k-means axiales" en dessinant sur la carte les liaisons inter-classes avec un trait d'intensité variable, selon la force de leurs connexions [Polanco et François, 2000].

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 d'interpréter la carte présentée en 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.

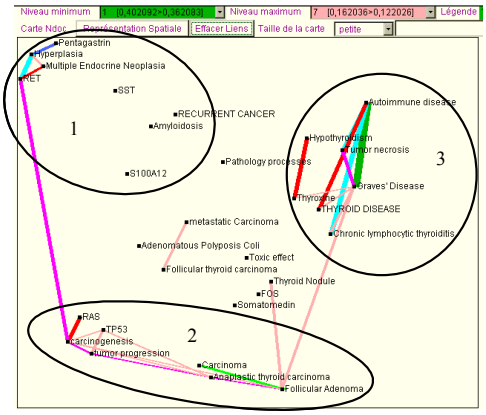


FIG. 2 - *Carte globale obtenue par classification Neurodoc réalisée à partir de l'index « Protéines+Maladies »*

Groupe 1 : Il 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 rassemble 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 relatives au gène RET : les cancers papillaires de la thyroïde (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.

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

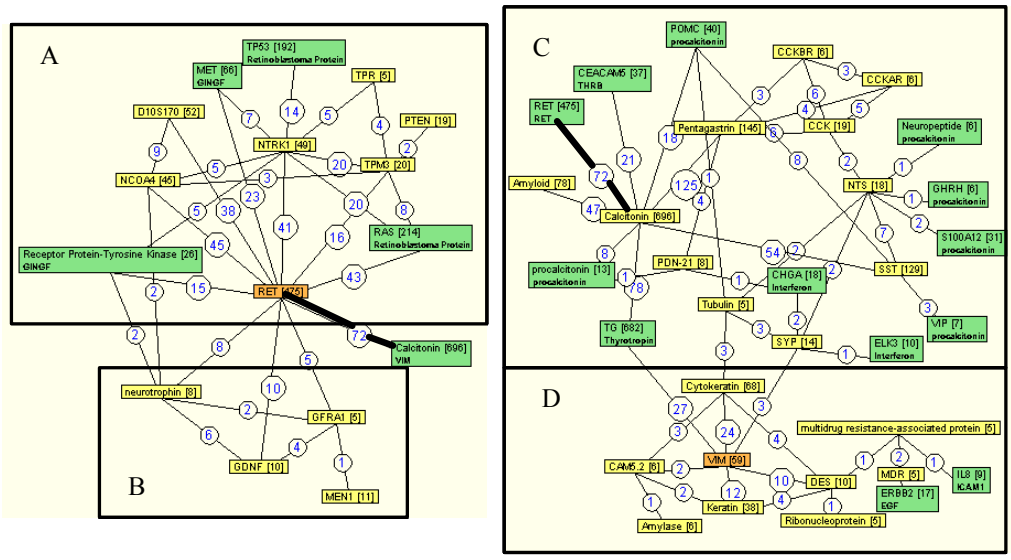
4.2. Sdoc : classification et réseaux lexicaux

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* [Callon et al., 1986] dont les valeurs varient entre 0 et 1. A partir de ce réseau, 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 et 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.

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

La classification Neurodoc met en évidence l'implication du gène RET dans les deux pathologies cancéreuses de la thyroïde : les PTC et MTC. Le sous-réseau associé à ce gène, obtenu avec la classification Sdoc (figure 3), se répartit entre deux classes (« RET » et « VIM ») reliées entre elles par une association externe (« RET » - « Calcitonine ») et soulignée d'un trait plus épais sur cette figure. La classe « RET » est formée de deux sous-ensembles distincts montrant deux processus biologiques dans lesquels ce gène est impliqué : les mécanismes moléculaires des PTC (sous-réseau A), et la régulation neuroendocrine par des voies de signalisation intercellulaire (facteurs de croissance neuronaux et leurs récepteurs) (sous-réseau B). Le réseau de la classe « VIM » se divise en deux sous-réseaux C et D, chacun étant impliqué dans une pathologie tumorale de la thyroïde.



Classe «RET» Classe «VIM»

FIG. 3 - *Réseau des noms de gènes et protéines des classe « RET » et « VIM » obtenues avec la classification Sdoc.*

Sous-réseau A : Ce sous-réseau associant RET à NCOA4, NTRK1, PTEN, D10S170, TPM3 et TPR, renseigne sur les principales interactions entre protéines ou gènes décrites dans la cadre de l'oncogenèse 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 (figure 2) 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 voies de signalisation intercellulaire.
Sous-réseau C : Ce sous-réseau 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 MTC.
Sous-réseau D : Ce sous-réseau autour du gène VIM concerne également les PTC. Les gènes associés à ce sous-réseau se retrouvent dans la description des classes situées sur la partie droite de la carte Neurodoc avec entre autres les classes « Anaplastic thyroid carcinoma » et « Follicular Adenoma » du groupe 2.
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.

4.3. Quelle démarche pour une analyse experte

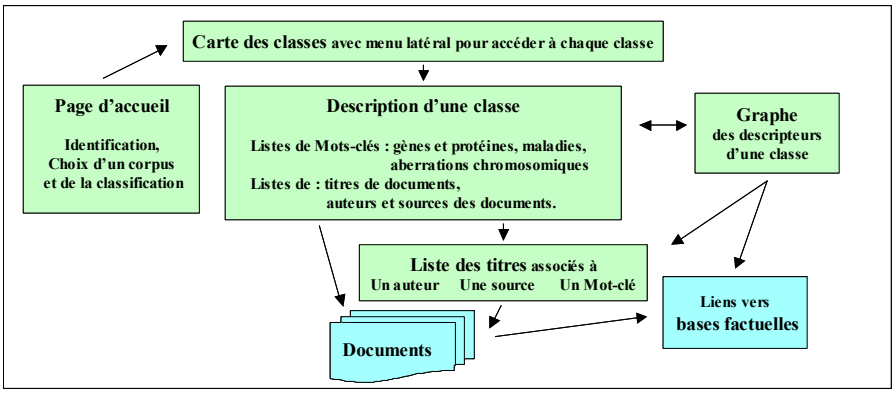


FIG4 – *Schéma de navigation simplifié dans le serveur VISA.*

Le processus de fouille décrit ici est un processus interactif et itératif. En effet, une validation humaine est nécessaire pour les phases d'indexation et de classification nécessitant une collaboration entre chercheurs des sciences de l'information et biologistes (figure 1). La validation de l'indexation consiste dans la détection des bruits et silences de l'indexation qui peuvent trouver leur cause dans la couverture et la qualité des ressources terminologiques mais aussi dans l'adéquation des outils avec le type de vocabulaire traité. De plus, les classifications nécessitent de tester plusieurs paramétrages avant de sélectionner le résultat qui sera présenté aux experts. Cette phase de classification peut également faire émerger les bruits ou silences de l'indexation nécessitant de revenir à cette phase d'indexation, soit avec de meilleurs paramétrages linguistiques, soit par des développements spécifiques pour traiter un problème non pris en compte avec les outils existants.

L'analyse des classifications est réalisée à l'aide d'une interface Web (serveur VISA) qui permet d'accéder non seulement aux cartes et graphes présentés dans les figures 2 et 3, mais également à la description complète de chaque classe incluant la liste des documents qui leur sont affectés, et les notices bibliographiques associées (figure 4). L'accès aux résumés est nécessaire pour valider les relations et proximités entre concepts calculées statistiquement et analyser leur contexte. L'accès aux bases factuelles décrivant et définissant les gènes et protéines rencontrés dans la littérature est également une aide très précieuse à l'interprétation des classes et réseaux. L'utilisation de plusieurs méthodes de classification sur des index différents nécessite d'étudier la complémentarité des informations recueillies par ces deux modes. La classification Neurodoc a permis de retrouver les associations principales entre les gènes et les pathologies de la thyroïde dans les listes de mots-clés décrivant les classes. Les concepts des différents sous-réseaux obtenus par la classification Sdoc ont également été retrouvés dans la description des classes Neurodoc permettant de confirmer les regroupements opérés et de les compléter par la description des réseaux de gènes.

L'analyse de ce type de résultat nécessite une première phase de détection des structures obtenues avec les différentes méthodes, complétées par une expertise permettant de valider ces dernières avec l'aide des connaissances de l'expert et des informations disponibles au travers de l'interface utilisateur.

5. Conclusion

Notre travail a porté sur la mise en évidence des relations entre gènes (ou protéines) et les pathologies tumorales de la thyroïde, par des méthodes terminologiques d'indexation et des méthodes d'analyse de données. L'utilisation de vastes ressources hétérogènes provenant de deux bases lexicales a permis d'obtenir pour chaque notice bibliographique un index contrôlé et structuré, destinés à des classifications ciblées, où chaque terme appartient à une ou plusieurs catégories sémantiques. Deux de ces catégories ont retenu notre intérêt : les maladies et les gènes. Chaque classe est donc annotée par les maladies et gènes cités dans les articles où ils sont présents. La classification Neurodoc met 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 les pathologies liées à des dysfonctionnements du système immunitaire. Ces classes confirment 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 montre le rôle de RET dans les deux principaux cancers de la thyroïde tout en précisant les relations qu'il établit avec les autres gènes. Les informations et relations mises en évidence, connues des experts, n'apportent pas, d'éléments nouveaux et/ou des rapprochements inattendus. Mais, notre ambition, dans le projet, était la mise en place d'une méthodologie d'analyse facilitant le travail des experts. 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.

Les auteurs remercient également Patrick Millan pour les développements informatiques apportés au serveur de visualisation des résultats (VISA).

Références

[Aronson, 2001] AR. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, Proc AMIA Symp. 2001;:17-21.

[Bessières et al., 2001] 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.

[Blaschke et Valencia 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, et al., 1986] M. Callon, J. Law, A. Rip., (editors) Mapping the Dynamics of Science and Technology. London, MacMillan Press, 1986.

[Cohen, 2002] 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.

[Daille, 2001] 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, 1995] 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.

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

[Jacquemin et Tzoukermann , 1999] 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.

[Jacquemin, 2002] 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.

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

[Lelu et François, 1992] 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/.

[Mac Queen, 1967] 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.

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

[Pillet, 2000] 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

[Poibeau, 2001] 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.

[Polanco et François, 2000] 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.

[Royauté, 1999] 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.

[Royauté, 2001] J. Royauté, C. François et D. Besagni., 2001- Apport d'une méthodologie de recherche de termes en corpus dans un 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.

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

[Schwartz et Hearst, 2003] 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

[Schmid, 1994] 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). (http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML)

Summary

Genomic and proteomic relationships in thyroid cancers are highlighted through the analysis of a large corpus of abstracts extracted from the Medline bibliographic database. A multidisciplinary approach (biologists, clinicians, linguists and researchers in information science) has enabled an automatic indexation and analysis of this corpus. The 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, takes into account the specificity of terms: biochemical nomenclatura, gene acronyms, chromosomic aberrations, or linguistic variants of terms. Two complementary classification methods reveal a dense lexical network of coocurrent genes which is linked to three main thyroid pathologies: medullary carcinomas (MTC), papillary carcinomas (PTC) and immune system dysfunction. In addition, the enhancement of the interactive visualisation tools of INIST's VISA server facilitate document access and navigation.