

# Relations entre gènes impliqués dans les cancers de la thyroïde<sup>1</sup>

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 cooccurents 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

---

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