L'induction de graphes dans l'étude du complexe *Mycobacterium tuberculosis.*

Georges Valétudie*, Séverine Ferdinand**
Nalin Rastogi**, Christophe Sola**

(*) Université des Antilles-Guyane, UFR Sciences Exactes et Naturelles
Laboratoire GRIMAAG EA 3590, Campus de Fouillole
BP 97110 Pointe-à-Pitre Guadeloupe
georges.valetudie@univ-ag.fr

(**) Unité de la Tuberculose et des Mycobactéries, Institut Pasteur de Guadeloupe
{sferdinand, nrastogi, csola}@pasteur-guadeloupe.fr

Résumé. Du fait de l'évolution parallèle des méthodes d'identification moléculaire des génomes mycobactériens et de leur structure, la reconnaissance et la détection des gènes deviennent des enjeux majeurs dans le domaine de la santé publique. Minimiser les coûts et le nombre de tests d'analyse et d'identification par des techniques assurant un résultat satisfaisant s'apparente en informatique à la recherche d'attributs pertinents capables de discriminer efficacement les individus au sein de la structure étudiée. Dans cet article, nous axerons notre recherche sur l'étude de la contribution de l'induction supervisée dans le classement de données issues de la tuberculose par une approche exploitant conjointement spoligotypes et MIRU-VNTR.

1 Introduction

L'étude du complexe ***Mycobacterium Tuberculosis*** constitue un enjeu majeur en matière de santé publique dans la lutte contre la tuberculose. Les chercheurs disposent de techniques éprouvées mais souvent coûteuses ou longues à mettre en œuvre. Les données de génotypage sont obtenues par des expériences de laboratoire faites sur de l'ADN. Dans notre cas, les données sont obtenues en appliquant la technique dite de **spoligotyping** (Kamerbeek *et al.*, 1997) sur de l'ADN préparé au préalable après repiquage de souches de *Mycobacterium Tuberculosis*, principalement isolées à partir des prélèvements reçus par l'Institut Pasteur, en provenance de la Guadeloupe. Par ailleurs, compte tenu des activités de référence de l'Institut Pasteur, des souches arrivent pour identification de Martinique et de Guyane française. De plus, des ADN sont envoyés par les centres GHESKIO à Port-au-Prince (Groupement Haïtien d'Etude du Sarcome de Kaposi), situés en Haïti, avec lesquels l'Institut Pasteur a entamé une collaboration depuis 1999. C'est dans le cadre d'une collaboration scientifique que ces données ont été mises à notre disposition.

Notre but est d'extraire des connaissances par le biais de modèles adaptés à ce type de données séquentielles, à partir d'une base de données conséquente et en constante augmentation. Cela consistera à chercher les séquences les plus discriminantes de classes d'individus définies *a priori* par les experts du domaine et à automatiser par des règles de