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

connaissances les procédures de traitement de séquence d'ADN, à la fois coûteuses et longues.

Une première étude consistant à faire appel aux méthodes associant des indices de dissimilarité a été menée dans (Valetudie *et al.*, 2001) et a mis en évidence la difficulté de classifier les données de spoligotypage. Des résultats certes intéressants sont obtenus par les méthodes axées sur les mesures de proximité, mais l'objectif y est « seulement » d'optimiser les taux de succès en généralisation. Rappelons que nous travaillons dans le domaine de la santé publique et que notre but est de repérer certaines caractéristiques qui permettraient d'établir la nature d'une souche à partir de celles déjà représentées dans la base. L'objectif premier est donc de construire un *classifieur* capable d'identifier la classe d'appartenance d'un nouvel individu. D'autre part, une contrainte sous-jacente à notre problème, est de produire un modèle intelligible, capable d'être interprété par les experts du domaine, afin de valider sous forme de règles leurs connaissances. Pour toutes ces raisons, nous nous sommes orientés vers les arbres de décision.

Nous commencerons notre discours en présentant la technique du spoligotyping et l'apport de la sélection de prototype dans l'amélioration de la qualité de l'échantillon d'étude. Après quelques rappels succincts sur la méthode des arbres de décision, nous évaluerons l'impact de son application dans le cadre du processus de classement du complexe *Mycobacterium tuberculosis*. Nous confronterons les résultats obtenus à partir des spoligotypes avec ceux obtenus sur la base de l'association de ces derniers avec les données de type MIRU-VNTR (**M**ycobacterial **I**nterspersed **R**epetitive **U**nits-**V**ariable **N**umber of **T**andem DNA **R**epeats). Enfin, nous préciserons les nouvelles perspectives vers lesquelles nous nous orientons.

2 Le spoligotyping et la sélection de prototypes

La plus grosse partie des données disponibles sont celles issues du spoligotypage qui reste une technique très utilisée (Sola *et al.*, 1999 ; Legrand *et al.*, 2001) et de ce fait incontournable. De la méthode dite de **spoligotyping** pour "**sp**acer **oligo**nucléotide **typing**" (Kamerbeek *et al*, 1997), découle le nom de **spoligotype** donné aux représentations binaires des composantes de la base. Chacun de ces bits constituent les **espaceurs** (spacers) (voir un exemple sur la Figure 1). Ces données peuvent se présenter sous forme binaire, octale ou hexadécimale établie selon les règles de standardisation publiées dans (Dale *et al*, 2001). La base dénommée **SpolDB3**, que nous exploitons dans la partie expérimentale, est à dimension internationale et comporte actuellement **11708** *spoligotypes* de souches issues de **90** pays. Pour l'heure, **813** shared-types **(ST)**, ou familles, ont été définis, **1300** souches restant actuellement orphelines. Cette base de données fait l'objet d'une description complète dans (Filliol *et al*, 2003).

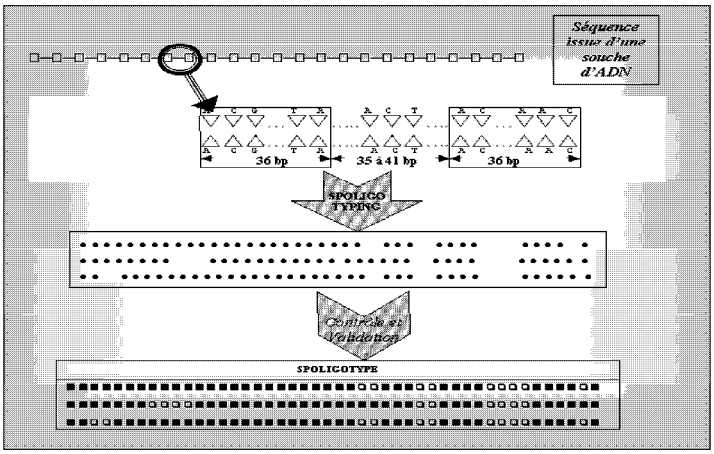

FIG. 1. *Cheminement des données de la souche au spoligotype : celles-ci sont retranscrites en carrés (noir ou blanc) qui représentent les espaceurs*

Soucieux d'éloigner les données bruitées (non informatives, aberrantes ou redondantes), nous avons étudié différentes méthodes de prétraitement dont le comparatif est proposé dans (Nock *et al.*, 2001). Notons que le pré-traitement des données, par la sélection de prototypes (PS), constitue une approche efficiente permettant d'accroître les performances de classement, tout en réduisant la complexité algorithmique des modèles. Si la plupart des méthodes PS privilégient le taux de succès comme principal critère à optimiser sur l'échantillon d'apprentissage, d'autres critères tels que la réduction des contraintes de stockage des données, la résistance au bruit, et la vitesse d'apprentissage sont tout aussi importants pour juger de la qualité d'une méthode. L'étude comparative menée dans (Nock *et al.*, 2001) entre les méthodes *Condensed Nearest Neighbor Rule* (Hart, 1968*), Consensus Filter* (*CF*) (Brodley *et al*, 1996), *Monte-Carlo sampling* (Skalak, 1994), *RT3* (Wilson *et al.*,1997), *PSRCG* (Sebban *et al.*, 2000), et *PSBOOST* fait ressortir toute la difficulté de réduire la taille de l'échantillon tout en conservant une pertinence lors de la généralisation. Ainsi, la méthode *CF* apporte de bons résultats pour la généralisation mais ne résout pas réellement le problème de la réduction de données. *A contrario*, la méthode *RT3* réduit de façon drastique la taille de l'échantillon, mais ceci au détriment de la généralisation. Il ressort donc que l'apport de la technique du *boosting* (Freund *et al. 1996*), exploitée dans l'algorithme *PSBOOST* que nous utilisons dans notre étude, donne le meilleur compromis entre les différents critères optimisés.

3 Une méthode d'analyse basée sur les arbres de décision

Parmi les différentes catégories de méthodes d'apprentissage (modèles connexionnistes, modèles statistiques, modèles symboliques, etc…), les arbres de décision se particularisent par la possibilité d'édition de règles aisément interprétables pour la plupart des problèmes d'apprentissage supervisé. Particulièrement bien adaptés à des données symboliques, leur construction nécessite des techniques de discrétisation dès lors qu'on traite des données

numériques continues. La perte d'information est souvent largement compensée par le niveau informationnel des modèles induits. Le principe général consiste à opérer une succession de partitions sur une population, par l'optimisation d'un critère de qualité. Ce peut-être une entropie quadratique, l'indice de GINI dont la formule[1] est :

$$I_{G(S)} = \sum_{j=1}^{k} \frac{n_{.j}}{n} \sum_{i=1}^{m} \frac{n_{ij}}{n_{.j}} \left(1 - \frac{n_{ij}}{n.j}\right)$$

ou encore l'entropie de Shannon (Shannon *et al,* 1949) :

$$I_{s(S)} = \sum_{j=1}^{k} \frac{n_{.j}}{n} \left(- \sum_{i=1}^{m} \frac{n_{ij}}{n_{.j}} \log_2 (\frac{n_{ij}}{n_{.j}}) \right)$$

voire le Gain Ratio :

$$I_{GR(S)} = \frac{- \sum_{i=1}^{m} \frac{n_{i.}}{n} \log_2 (\frac{n_{i.}}{n}) + \sum_{j=1}^{k} \frac{n_{.j}}{n} \sum_{i=1}^{m} \frac{n_{ij}}{n_{.j}} \log_2 (\frac{n_{ij}}{n_{.j}})}{- \sum_{j=1}^{k} \frac{n_{.j}}{n} \log_2 (\frac{n_{.j}}{n})}$$

Le critère retenu dépend en fait de la méthode employée pour la construction de l'arbre. La méthode *C4.5*, proposée dans (Quinlan, 1992) figure parmi les références en la matière. Elle fait appel au gain ratio comme critère de mesure d'incertitude. La méthode *CART* (Breiman *et al.*, 1984), qui fait aussi figure de référence, emploie l'indice de Gini comme critère d'optimisation. La méthode *ID3,* également proposée par *Quinlan*, utilise *le gain d'information***,** une mesure d'incertitude basée sur l'Entropie de Shannon. Enfin, la méthode *SIPINA* (Zighed *et al.*, 1992) qui constitue une forme de généralisation des méthodes existantes, a pour particularité de proposer la construction de graphes non-arborescents.

Chacun des chemins du graphe construit correspond à une règle de production exprimable selon le formalisme de la logique propositionnelle :

Si < *Prémisses(s)* > **Alors** => < *Conclusion* > [*Taux de confiance*]

- *Prémisses :* représente la conjonction de conditions portant sur les valeurs que doivent vérifier les attributs considérés,

- *Conclusion :* illustre la décision associée aux prémisses compte tenu de la valeur du critère de qualité de partition relativement au seuil de décision fixé,

- *Taux de confiance :* précise le degré de confiance à accorder à la règle.

C'est avec l'ensemble de ces règles que nous avons construit le modèle de prédiction recherché. La qualité du modèle et la confiance qu'on lui accordera est intrinsèquement liée à celle des règles dont il découle.

Des indicateurs nous permettent de mesurer les taux de reconnaissance ou de rejet, ainsi que le taux de confiance à accorder à la règle induite. Ces indicateurs prennent tout leur sens dans

[1] **S** : partition étudiée, **m** : nombre de classes, **n** : taille de l'échantillon, **k** : nombre d'éléments de la partition, $\mathbf{n_{ij}}$: nombre d'éléments de classe i dans la partition j, $\mathbf{n_{i.}}$: nombre d'éléments de classe i, $\mathbf{n_{.j}}$: effectif de la partition j.

notre domaine d'application dans la mesure où on préfèrera rejeter une hypothèse plutôt que d'accorder sa confiance à une conclusion hasardeuse. Les indicateurs utilisés portent sur :

- l'*effectif,* qui fournit un premier indice sur l'influence de la règle,

- la *précision de la règle,* qui complète l'information précédente en indiquant la proportion d'individus correctement classés parmi ceux concernés par la règle,

- la *pertinence de la règle,* qui mesure l'intérêt de la règle au vu du nombre d'exemples mal-classés.

4 Résultats et expérimentation sur nos données

Nous prenons pour le complexe *Mycobaterium Tuberculosis* une présentation en neuf classes, *Afri, T, Beijing, EAI, Haarlem, LAM, CAS, X et Bovis*, dont les caractéristiques ont été établies dans (Sebban *et al.*, 2001) puis affinées et complétées dans (Filliol *et al.,* 2003). Notre analyse poursuit les travaux enclenchés, sur la base de données récentes et non encore exploitées, certaines familles s'étant enrichies en effectif et en qualité.

Ces classes ont été établies *a priori* par les experts du domaine, qui ont mis en exergue les relations existant entre les spoligotypes de même classe. Ainsi, après expertise humaine, on a constaté que les Beijing se caractérisent par l'absence des spacers 1 à 24, Les T par l'absence de 33 à 36, les EAI par l'absence de 29 à 32 et 34 et la présence de 33, les LAM par l'absence de 21 à 24 et de 33 à 36, les CAS par l'absence de 4 à 7, les X par l'absence de 18 et 33 à 36, les Bovis par l'absence de 39 à 43. Notre démarche consistant à construire des arbres alors qu'il y a déjà des règles d'experts s'explique par trois arguments essentiels :

- vérifier que les règles de l'expert sont bonnes en terme de performances,

- essayer de simplifier ces règles qui parfois utilisent trop d'espaceurs pour être facilement exploitables,

- si le taux de succès n'est pas bon avec les arbres, il s'agira peut-être de réétiqueter certains exemples et voir ainsi émerger d'autres classes.

Notre fichier de données se répartit selon les effectifs suivants:

	Sans PSBOOT		Avec PSBOOT	
Classe	Effectif	Taux	Effectif	Taux
AFRI	42	5 %	14	5 %
BOVIS	56	7 %	19	7 %
BEIJING	11	1 %	4	1 %
CAS	29	4 %	11	4 %
EAI	99	13 %	36	13 %
HAARLEM	94	12 %	35	13 %
LAM	176	22 %	64	23 %
T	226	29 %	75	27 %
X	52	7 %	17	6%
Total	785	100 %	275	100 %

TAB 1 – *Composition du fichier de spoligotypes Dat1*

Classement du complexe. *Mycobacterium tuberculosis*

La plate-forme *SIPINA for Windows*© (Rakotomalala *et al,* 2000) qui inclut les différentes méthodes utilisées, a été exploitée pour établir les graphes et les règles de décision correspondantes. Nous avons procédé à une phase d'apprentissage puis de validation dont les résultats sont consignés dans le tableau 2.

Pour le choix de la méthode, nous avons comparé le taux de classement sans tenir compte des cas rejetés (colonne 1) puis mesuré l'impact engendré par la prise en compte des individus rejetés (colonne 2) en ramenant le taux de rejet au taux d'observations bien classées. Ce comparatif a été modulé par le nombre de niveaux de l'arbre qu'il convenait de minimiser. Malgré un bon résultat obtenu sans prise en compte des non-classés (97 %), les performances de CHAID s'écroulent dès lors qu'ils sont pris en compte. La méthode C4.5 présente des performances similaires à ID3 mais semble globalement la plus satisfaisante en terme de qualité de classement, avec une meilleure précision des résultats compte tenu des individus non-classés, dont l'effectif est ici le plus faible, et du nombre de niveaux tout à fait raisonnable qui laisse présager des règles de décision assez concises. C'est donc cette méthode que nous choisissons pour l' étape suivante du processus. Nous obtenons les règles suivantes ainsi que leur évaluation, la valeur **Y** marquant la présence du spacer, la valeur **N** son absence.

Méthode appliquée	Taux d'observations correctement classée (%)		Nombre de non classés	Nombre de niveaux de l'arbre
	Sur l'effectif sans les non-classés	Sur l'effectif avec les non-classés		
ID3	96	89	22	7
C 4.5	95	90	14	7
Cart [gini]	93	80	31	6
Sipina	93	82	33	6
Chaid	97	67	85	6

TAB 2 – *Pertinence des méthodes lors du classement*

18 = Y and 22 = Y and 31 = Y and 34 = N and 33 = N and 36 = N then Classe = T

22 = N and 21=N and 24=N and 34=N and 36=N then Classe = LAM

12 = N and 33 = N and 36 = Y and 39 = Y then Classe = BEIJING

34 = N and 31 = N and 32 = Y and 33 = N and 36 = N and 22 = Y then Classe = HAARLEM

34 = N and 36 = N and 31 = Y and 18 = N and 22 = Y and 33 = **N** then Classe = X

34 = N and 33=Y then Classe=EAI

34 = N and 36 = Y and 33 = N then Classe = CAS

43 = Y and 34 = Y then Classe = AFRI

34 = Y and 43 = N then Classe = BOVIS

Ces règles sont décrites selon les principes qui ont été définis auparavant. Pour exemple, la première règle traduit l'assertion suivante : « Si dans le spoligotype étudié, les spacers 18, 22 et 31 sont présents et les spacers 33, 34 et 36 sont absents, alors la classe associée est la classe T. ».

La qualité associée à chacune de ces règles est précisée dans le tableau 3 :

Classe	Nb espaceurs	Pertinence de la règle (%)	Individus bien classés (%)
Afri	2	99.99	100
Bovis	2	99.99	83
Haarlem	6	100	100
Eai	2	100	97
Cas	3	99.72	71
Beijing	4	99.96	100
T	6	100	100
X	6	99.95	76
Lam	5	100	98

TAB 3 – *Qualité des règles avec C4.5*

Classe	Effectif	Taux
Afri	22	7 %
Bovis	10	3 %
Beijing	7	2 %
Cas	10	3 %
Eai	46	14 %
Haarlem	53	16 %
Lam	61	18 %
T	99	30 %
X	25	8 %
Total	333	100 %

TAB 4 – *fichier Dat2*

Suivant ces résultats, les règles s'avèrent globalement pertinentes avec des taux de reconnaissance avoisinant le plus souvent 100 % des individus. Seuls CAS et X ont un taux inférieur à 80 % du fait d'erreurs d'étiquetage selon l'expert. Cela conforte le choix de la méthode C4.5.

L'étape suivante de notre application consiste à valider le classifieur obtenu en mesurant la capacité de nos règles à reconnaître le profil de 333 individus d'un échantillon test. Ces derniers sont répartis selon les neuf classes du tableau 4.

Le taux de classement correct obtenu à l'issue de cette phase est de l'ordre de 88 %.

Les résultats détaillés sont consignés dans le tableau 5.

Au-delà de ces résultats qui confirment assez bien la qualité du classifieur, il ne faut cependant pas négliger les 12 % d'individus qui restent encore mal-classés. A bien observer les résultats portant notamment sur nos règles, on constate que l'imprécision relevée provient principalement des classes T et BEIJING.

C'est donc pour palier à ces difficultés que nous avons pris en compte la possibilité d'exploiter d'autres types de données, en l'occurrence les données de MIRU-VNTR. En effet, des travaux antérieurs (Supply *et al.*, 2001) ont souligné l'intérêt de ces marqueurs en épidémiologie génétique qui restent inexploités pour l'instant dans notre domaine.

Classement du complexe. M. Tuberculosis

Classe	Spoligos bien classés		Spoligos mal classés		Spoligos rejetés	
	Effectif	Taux	Effectif	Taux	Effectif	Taux
AFRI	17	6 %	5	14 %	0	0 %
BEIJING	0	0 %	7	19 %	0	0 %
BOVIS	9	3 %	1	3 %	0	0 %
CAS	10	3 %	0	0 %	0	0 %
EAI	39	13 %	7	19 %	0	0 %
HAARLEM	48	16 %	5	14 %	0	0 %
LAM	61	21 %	0	0 %	0	0 %
T	89	30 %	10	27 %	0	0 %
X	23	8 %	2	5 %	0	0 %
Total	296	100 %	37	100 %		0 %
% de l' effectif total	89 %		11 %		0 %	

TAB 5 – *Résultats après validation sur le fichier de spoligotypes Dat2*

5 Apport des données de MIRU-VNTR dans notre analyse

Les données de MIRU-VNTR se présentent sous une forme numérique composée de 12 chiffres, chaque valeur indiquant le nombre de répétitions sur chacun des 12 loci (régions) répétés. Ces répétitions d'ADN bactérien sont analogues aux répétitions « minisatellites » rencontrées dans le génome humain (Frothingham *et al.*, 1998). Ces données ont été étiquetées par l'expert en tenant compte des classes établies à l'aide des données de spoligotypage.

Notre fichier de données se répartit selon les effectifs suivants :

Classe	Effectif	Taux	Classe	Effectif	Taux
AFRI	4	3 %	HAARLEM	13	9 %
BOVIS	4	3 %	LAM	25	17 %
BEIJING	4	3 %	T	24	16 %
CAS	0	0 %	X	56	38 %
EAI	16	11 %	Total	146	100 %

TAB 6 – *Composition du fichier MIRU-VNTR Dat3*

Les résultats que nous présentons sont pour l'heure des résultats préliminaires. La méthode ID3 fournit le meilleur taux lorsque les non-classés ne sont pas pris en

compte. De plus, le nombre de mal classés est plus faible (8 contre 12) ce qui est important dans ce domaine d'application (cf tableau 7). Voici les règles et mesures de qualité associées :

MIRU10 < 5 and MIRU24 < 2 and MIRU31 > 4 then Classe = BEIJING

MIRU23 < 5 and MIRU24 < 2 then Classe = HAARLEM

MIRU10 < 4 and MIRU40 > 3 and MIRU24 < 2 and MIRU23=5 then Classe = T

MIRU10 > 3 and MIRU40 < 3 and MIRU24 < 2 and MIRU23>5 then Classe = LAM

MIRU24 < 2 and MIRU23 = 5 and MIRU10 = 4 then Classe = X

MIRU24 > 1 and MIRU26 < 3 then Classe = EAI

MIRU24 > 2 and MIRU26 > 2 then Classe = BOVIS

Méthode appliquée	Taux d'observations correctement classée (%)		Nombre de non classés	Nombre de niveaux de l'arbre
	Sur l'effectif sans les non-classés	Sur l'effectif avec les non-classés		
ID3	90	51	63	7
C 4.5	88	65	39	8
Cart [gini]	88	19	87	3
Sipina	85	52	56	5
Chaid	80	19	111	5

TAB 7 – *Pertinence des méthodes à l'issue de la phase d'apprentissage*

Classe	Nombre d'attributs	Pertinence de la règle	Individus bien classés
Afri	Indéterminé		
Bovis	2	90,76 %	57 %
Haarlem	2	97,11 %	64 %
Eai	2	99,99 %	100 %
Beijing	3	99.96 %	100 %
T	4	86,94 %	56 %
X	3	99.99 %	89 %
Lam	4	97,94 %	86 %

TAB 8 – *Qualité des règles avec ID3*

Méthode	Qualité moyenne	Biais (%)
C4.5	67 %	17
Id3	68 %	20
Sipina	68 %	20
Méthode	**Nombre moyen de règles**	**Biais**
C4.5	9	1.05
Id3	11.89	0.87
Sipina	6	0

TAB 9 – *Cross-Validation stratifiée avec 9 sous-groupes*

Classement du complexe. M. Tuberculosis

Ces données permettent d'obtenir des résultats encourageants en phase d'apprentissage, mais elles sont encore insuffisantes pour valider efficacement le modèle défini. Certaines classes en effet sont indéterminées et réclament un affinage complémentaire des règles pour mieux les définir. Les règles établies ne concernent ici que 8 familles sur les 9 définies, la famille CAS n'étant pas représentée par défaut d'effectif. Par ailleurs, l'étiquetage de ces données établi à partir des spoligotypes est probablement à l'origine de certaines incohérences dans les observations. Nous travaillons pour l'heure sur une approche non supervisée afin d'établir les classes potentielles issues de MIRU-VNTR et déterminer leur corrélation avec celles issues des spoligotypes ainsi que la nature et la qualité des règles établies dans l'analyse de leur contribution respective dans la reconnaissance des spoligotypes.

6 Conclusion et perspectives

Nous avons étudié l'apport de la sélection de prototype et de la méthode des arbres de décision dans l'étude des spoligotypes. Les résultats obtenus par induction supervisée sont encourageants. Le modèle de règles inféré permet en effet de classer les spoligotypes avec un taux de succès intéressant pour les biologistes de l'Institut Pasteur. Cependant certaines familles présentent des difficultés dans la phase de reconnaissance du fait de la grande diversité de leurs caractéristiques. Les règles intelligibles et pertinentes obtenues sont appréciées pour leur simplicité qui recadre l'approche visuelle menée jusqu'alors. L'impact méthodologique et financier qu'engendre à court terme une analyse aussi rapide et simple est omniprésent, notamment lorsqu'un bon pouvoir prédictif est couplé à une petite quantité de données issue du prototypage. Nous étudions l'approche présentée par (Borgi *et al.*, 1999) dans le but de prendre en compte les corrélations potentielles inter-attributs. Les résultats préliminaires issus des MIRU-VNTR nous autorisent à nous impliquer plus dans cette démarche.

Références

Borgi A., Akdad A., (1999). *Induction supervisée de règles : le système SUCRAGE* CAP'99 : Conférence d'Apprentissage Plate-forme, AFIA.

Breiman I., Friedman J. H., Olshen R. A., Stone C. J. (1994). *Classification and regression tree*, Chapman and Hall.

Brodley C.E., Frield M. (1996). *Identifying and eliminating mislabeled training instances,* Proc. 13th National Conference on Artificial Intelligence

Dale J. W., Brittain D., Cataldi A., Cousins D., Crawford J. T., Driscoll J., Heersma H., Lillebaek T., Quitugua T., Rastogi N., Skuce R., Sola C., van Soolingen D., Vincent V. (2001). *Spacer oligonucleotide typing of Mycobacterium tuberculosis complex : recommendations for standardized nomenclature.* Int. J. Tuberc. Lung. Dis. **5**(3), 216-219.

Filliol I. et al. (49 auteurs). (2003). A Snapshot of Moving and Expanding Clones of *Mycobacterium tuberculosis* and their Global Distribution Assessed by Spoligotyping In An International Study. *J. Clin. Microbiol.* in Press.

Freund Y., Schapire R., (1996). *Experiment with a new boosting algorithm* Proc. 14th International Conference on Machine Learning, 148-156

Frothingham R., Meeker-O'Connell W. A. (1998). Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. *Microbiol.* **144**, 1189-1196.

Hart P. (1968). *The Condensed Nearest Neighbor Rule*, IEE. Trans Info. Theory, 515-516

Kamerbeek J., Schouls L., Kolk A., van Agterveld M., van Soolingen D., Kuijper S., Bunschoten A., Molhuizen H., Shaw R., Goyal M., van Embden J. D. A. (1997). Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.* **35**, 907-914.

Legrand E., Filliol I., Sola C., Rastogi N. (2001). Use of Spoligotyping To Study the Evolution of the Direct Repeat Locus by IS6110 Transposition in Mycobacterium tuberculosis. *J Clin Microbiol* **39**(4), 1595-1599.

Nock, R., Sebban M. (2001). *Advances in adaptative prototype weighting and selection.* International Journal on Artificial Intelligences Tools. Vol. 10 n° 1-2. World Scientific Publishing Company

Quinland J. R. (1992). *C4.5: Program for Machine Learning*, Morgan & Kaufmann.

Rakotomala R., Zighed D.A. (2000*). Graphes d'induction,* Paris Hermes

Sebban M., Mokrousov I., Rastogi N., Sola C. (2002). A Data-mining approach to Spacer Oligonucleotide Typing of *Mycobacterium tuberculosis*. *Bioinformatics* **18**, 235-243.

Skalak D (1994) Prototype and feature selection by sampling and random mutation hill climbing algorithms, Proc. 11th International Conference on Machine Learning, 293-301.

Sola C., Devallois A., Horgen L., Maïsetti J., Filliol I., Legrand E., Rastogi N. (1999). Tuberculosis in the Caribbean: using spacer oligonucleotide typing to understand strain origin and transmission. *Emerg. Inf. Dis.* **5**, 404-414.

Supply P., Lesjean S., Savine E., Kremer K., van Soolingen D., Locht C. (2001). Automated high-throughput genotyping for study of global epidemiology of Mycobacterium tuberculosis based on mycobacterial interspersed repetitive units. *J Clin Microbiol* **39**, 3563-71.

Valetudie G., Filliol I., Rastogi N., Sebban M., Sola C. (2001).*Classification des bacilles tuberculeux par génotypage : nouvelles méthodologies bio-informatiques par recherche de nouveaux indices de dissimilarité et par procédure d'élagage*, VIIIéme congrès de la Société Francophone de Classification , Pointe-à-Pitre.

Wilson D., Martinez T. (1997). *Instance Pruning Technics.* Proc. 14th International Conference on Machine Learning

Zighed D. A., Auray J.P., Duru G. (1992). *SIPINA : Méthode et logiciel* (Lacassagne, E. A., Ed.).

Classement du complexe. M. Tuberculosis

Summary

Recent progress in the molecular identification of mycobacterial genes and their structure has led to new developments in the field of rapid mycobacterial detection and molecular epidemiology permitting a better control of tuberculosis, which remains a global health priority. A wider applicability of these methods depends on the development of cost-effective and reproducible tests needing least human intervention. In this context, knowledge-discovery using data-mining methods permits to highlight relevant attributes of a given population in such a way that a minimum number of markers is enough to discriminate a maximum number of individuals. This is particularly true for the spoligotyping as well as for MIRU-VNTR (Mycobacterial Interspersed Repetitive Units - Variable Number of DNA Tandem Repeats) typing method, both permitting to discriminate mycobacterial isolates based on their respective polymorphisms. In this paper, we describe the contribution of various induction algorithms to correctly classify tubercle bacilli among some predominant families (clades) using spoligotyping and MIRU-VNTR.