Fouille de données biologiques : vers une représentation booléenne des règles d'association.

Abdelhak MANSOUL, Baghdad ATMANI

Equipe de recherche « Simulation, Intégration et Fouille de données (SIF) »
Département Informatique, Faculté des Sciences, Université d'Oran
BP 1524, El M'Naouer, Es Senia, 31 000 Oran, Algérie
http://www.univ-oran.dz
mans_abdel@yahoo.fr, atmani.baghdad@gmail.com

Résumé. L'avènement des biotechnologies nouvelles a permis, au cours des dernières années, d'accumuler des données sur les génomes des agents pathogènes épidémiologiques. Par contre l'exploitation des données génomiques n'as pas suivi le rythme des découvertes, alors la fouille de données biologiques, particulièrement à caractère épidémiologique s'est imposée d'elle-même afin d'aider à trouver des éléments de réponse aux questions que se pose l'épidémiologiste concernant des pathologies particulières.
D'où, la problématique abordée par cette étude qui est la fouille de données biologiques du Mycobacterium Tuberculosis responsable de la tuberculose. Nous proposons un processus de fouille de données assez novateur pour générer des connaissances qui vont êtres profitables et exploitables à deux niveaux :

- Profitables au spécialiste du domaine, à travers l'extraction de motifs en particulier les règles d'association qui aident à mieux comprendre la pathologie.
- Ensuite, ces règles d'association extraites sont modélisées par le principe booléen adopté par la machine cellulaire CASI (Cellular Automaton for Symbolic Induction).

Le but de cette modélisation par le principe booléen étant de réduire la complexité de stockage et le temps de réponse.

Mots clés: Automate cellulaire, Fouille de données biologiques, Induction de règles, Motif, Itemset, Règle d'association, Mycobacterium Tuberculosis, Tuberculose, Epidémie, Génome, Biologie.

1 Introduction

La biotechnologie a permis, au cours des dernières années, d'améliorer les connaissances sur le génome des agents pathogènes épidémiologiques, et de développer des moyens de lutte efficace contre ces épidémies. Actuellement, des dizaines de génomes ont été révélés et ont permis de constituer des banques de données biologiques énormes. De ce fait, les quantités de données brutes disponibles sont déjà trop importantes pour pouvoir être analysées manuellement par les méthodes épidémiologiques de surveillance et d'analyse. Du fait de l'inefficacité de ces méthodes

due à la variété des données biologiques, et à la nature même des épidémies, une nouvelle approche est utilisée : c'est la fouille de données biologiques relatives aux épidémies [2], [6]. Cette fouille permet d'extraire des connaissances qui serviront à mieux connaître les agents pathogènes, interpréter au mieux les phénomènes biologiques liés à une épidémie particulière, et ainsi permettre la mise en œuvre de mesures de prévention et de lutte, par des traitements appropriés, des vaccinations, .etc.

Problématique

Sur un terrain, purement épidémiologique, il y a une pathologie qui continue à faire des ravages et se trouve classée deuxième en mortalité après le sida : c'est la Tuberculose. Elle est l'un des plus grands fléaux de l'humanité qui entraîna en l'an 2000 près de 10 millions de nouveaux cas et plus de trois millions de morts chaque année dans le monde [19].

En effet, cette maladie infectieuse est provoquée par la pénétration dans l'organisme d'une bactérie appelée Mycobacterium Tuberculosis. Dans la pratique, il existe un Complexe Tuberculosis dont le Mycobacterium Tuberculosis est l'agent typique responsable de la tuberculose humaine [5].

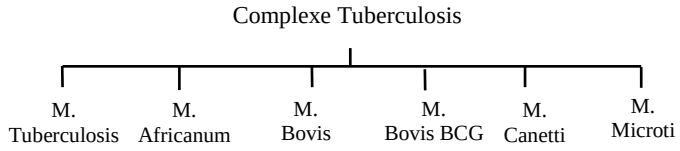


Fig. 1 Composition du Complexe Tuberculosis

En 1998, la première séquence complète du génome de Mt H37RV a été réalisée et a permis de dégager des caractéristiques propres aux mycobactéries dont les plus importantes sont les suivantes [5], [14], [18]:

- 51 % des gènes sont dupliqués;
- 10 % du génome code pour 2 familles de gènes qui codent eux même pour 2 protéines nommées PE et PPE;
- forte présence de séquences répétées d'ADN, en particulier une séquence nommée IS6110 (16 copies), riches en particularités sur le génome ;
- présence de 65 copies de MIRU (Mycobacterial Interspaced Repetitive Unit) ;
- présence de répétitions directes RD (appelées aussi régions de différences), ces séquences répétées sont riches en particularités sur le génome.

Tous ces éléments descriptifs de ce génome sont autant chacun un gisement qu'on exploite en fouille de données [8],[9],[20] afin d'essayer d'apporter des éléments de réponses à certains phénomènes liés au complexe Mycobacterium Tuberculosis, et trouver des solutions médicales afin de stopper la diffusion de la bactérie et par conséquent stopper l'épidémie par des vaccins, ou antibiotiques.

Donc, la problématique abordée dans ce papier, est la fouille de données biologiques se rapportant au Mycobactérium Tuberculosis à l'aide de tous les éléments d'informations cités auparavant à savoir : les gènes, les protéines, les RD, et les MIRU. Cette fouille se veut d'être une fouille de données hétérogènes.

Cette étude prendra en compte toutes les souches de la bactérie dont l'annotation a été complètement finie ou en projet de séquençage. Ce gisement de données sera plus conséquent s'il renfermera d'avantage de souches annotées, ce qui supposera par la suite, que toute souche nouvellement annotée, alimentera le processus de fouille de données envisagé.

2 Etat de l'art de la fouille de données biologiques

Depuis les premiers projets de séquençage des bactéries, les dispositifs expérimentaux tels que les séquenceurs automatiques, puces à ADN et autres, ont permis de constituer des bases de données de séquences de génomes complets. Il fallait donc exploiter ces données, identifier les gènes, les protéines qu'ils produisent, et identifier leurs fonctions, pour comprendre les mécanismes de la bactérie. De plus, la variété et la disponibilité des données biologiques (séquences ADN, Protéines, Puce ADN, ….) et par la même des banques de données biologiques (NCBI, EMBL, GenBank….), ont incité à les valoriser. Différents travaux promoteurs et novateurs, en fouilles de données biologiques ont été faits en se basant essentiellement sur les génomes et les cohortes [13], les uns ont un rapport direct avec l'épidémiologie alors que d'autres la touchent indirectement (génomique et protéomique), mais sont d'un grand apport pour la compréhension des maladies et par la même des phénomènes épidémiologiques. Nous présentons quelques uns, mais la liste n'est pas exhaustive.

En génomique : Pour identifier des gènes, comparer des séquences (rechercher des similarités) [7], rechercher et extraire des motifs fréquents [16], différentes approches ont été utilisées comme l'extraction des séquences répétées (n-grammes) [15] ou les modèles de Markov cachés (HMM) [11],[16]. Ces modèles (HMM) interviennent aussi dans l'analyse de séquences pour la détection de répétitions [10] ou encore la recherche de mots exceptionnels [11], la recherche de gènes candidats, la recherche de séquences exogènes ou hétérogènes pouvant renseigner sur un pathogène impliqué dans une maladie [12]. L'utilisation des modèles de Markov cachés a permis aussi d'identifier les séquences exogènes [12] susceptibles de contenir des gènes de virulence ou des gènes d'adaptation, ce genre de recherche améliore la compréhension du phénomène de résistance aux antibiotiques. Plusieurs travaux sur les séquences biologiques ont donné naissance à des programmes dont les plus connus et les plus utilisés par les biologistes sont les logiciels FASTA et BLAST [7].

En fouille de cohortes : Les cohortes ont souvent été utilisées dans le cas des épidémies [13], elles fournissent un tas de données médicales (cliniques, biologiques, et génétiques) sur des cas réels (sujets exposés, non exposés). Ces fouilles permettent de renseigner sur le rôle des facteurs génétiques et environnementaux d'une maladie. Les méthodes de classification, les règles d'association ont étés utilisées dans ce cas pour permettre la détection des relations gène-gène et gène-environnement [21].

3 Contribution

Nous nous proposons d'étudier les aspects physiologiques fondamentaux liés à la génomique de cette bactérie modèle, le Mycobacterium Tuberculosis. Ensuite étudier les outils de fouille de données pour l'extraction des connaissances et d'en dégager une approches expérimentable.

En premier, nous avons établi un état de l'art de la fouille des données avec certains détails d'une technique à une autre et qui ne sont pas forcement en rapport direct avec notre étude. Ensuite, une étude comparative des différents outils et méthodes existants a été faite afin d'utiliser la plus adaptée à l'objet de notre étude.

Deuxièmement, nous avons abordé l'étude de l'agent pathogène, afin de cerner la nature et le type de données biologiques qui nous intéressent et ainsi pouvoir localiser nos sources de données expérimentales.

Troisièmement, nous avons établi notre propre démarche expérimentale par un processus de fouille de données pour la génération des connaissances à partir de données biologiques. Ces connaissances vont êtres profitables et exploitables à deux niveaux :

1. En premier, profitables au spécialiste du domaine pour la compréhension de la pathologie.
2. En second, exploitables par la machine cellulaire CASI [4] pour l'inférence et la déduction.

Ce processus informatique ainsi établi procède en deux étapes, une fouille de données est faite dans un premier temps en utilisant l'algorithme Apriori et donnera des règles d'association, ensuite et dans un deuxième temps produire des règles booléennes inductives qui vont alimenter la base de connaissances de la machine cellulaire CASI, cette machine développée pour l'acquisition automatique incrémentale de connaissances par induction et la prédiction par déduction [4].

Ainsi, notre contribution a adopté la démarche suivante :

1. Etude et sélection des données biologiques relatives au Mycobactérium Tuberculosis ;
2. Extraction des motifs fréquents et des règles d'association respectives ;
3. Production des règles booléennes inductives pour la machine cellulaire CASI.

4 Conception du système

Notre système est composé de deux grands modules, le premier produit des règles d'association et les transmet au deuxième module (BRI) pour générer des règles booléennes basées sur le principe de la machine cellulaire CASI.

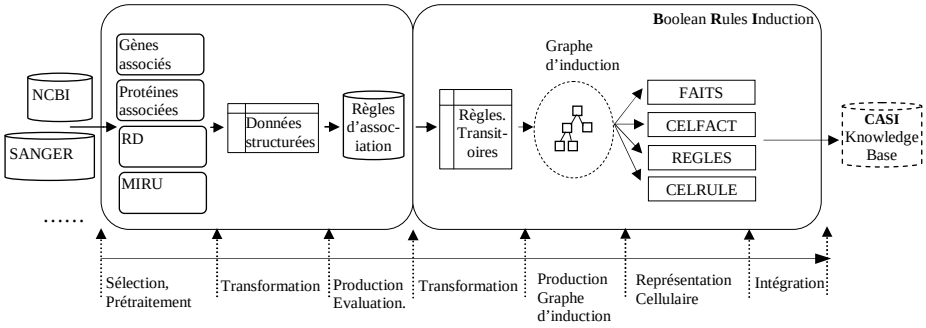


Fig.2: Fouille de données du complexe Mycobacterium Tuberculosis

4.1 Production des règles booléennes pour la machine cellulaire CASI

Les règles d'association produites sont transformées selon le principe suivant :

- Les items de Antécédent vont servir à constituer la Prémisse de la règle ;
- Les items de Conséquent vont servir à créer la Conclusion de la règle.

Cette transformation sert à produire des règles transitoires nécessaires à la production d'un graphe d'induction selon le principe suivant : Un sommet désigne un nœud sur lequel on fait un test, avec les résultats possibles binaires ou à valeurs multiples.
Ainsi le graphe d'induction permettra de produire les règles cellulaires sous la forme :

R_i : Si Prémisse$_i$ Alors Conclusion$_i$

Avec une représentation cellulaire selon le principe suivant :

- les items de Prémisse$_i$ et Conclusion$_i$ vont constituer les faits : FAITS.
- les R_i vont constituer les règles : REGLES.

Ces règles produites seront intégrées dans la base de connaissances de CASI pour exploitation en inférence.

4.2 La machine cellulaire CASI [1]

CASI (Induction Symbolique par Automate Cellulaire) est un automate cellulaire qui simule le principe de fonctionnement de base d'un Moteur d'Inférence en utilisant deux couches finies d'automates finis. La première couche, CELFACT, pour la base des faits et, la deuxième couche, CELRULE, pour la base de règles. Chaque cellule au temps t+1 ne dépend que de l'état des ses voisines et du sien au temps t. Dans chaque couche, le contenu d'une cellule détermine si et comment elle participe à chaque étape d'inférence : à chaque étape, une cellule peut être active (1) ou passive (0), c'est-à-dire participe ou non à l'inférence. Le principe adopté est simple :

- Toute cellule i de la première couche CELFACT est considérée comme fait établi si sa valeur est 1, sinon, elle est considérée comme fait à établir. Elle se présente sous trois états : état d'entrée (EF), état interne (IF) et état de sortie (SF).
- Toute cellule j de la deuxième couche CELRULE est considérée comme une règle candidate si sa valeur est 1, sinon, elle est considérée comme une règle qui ne doit pas participer à l'inférence. Elle se présente sous trois états : état d'entrée (ER), état interne (IR) et état de sortie (SR). Les matrices d'incidence R_E et R_S représentent la relation entrée/sortie des Faits et sont utilisées en chaînage avant et en chaînage arrière en inversant leur ordre.

La dynamique de l'automate cellulaire, pour simuler le fonctionnement d'un Moteur d'Inférence, utilise deux fonctions de transitions δ_{fact} et δ_{rule}, où δ_{fact} correspond à la phase d'évaluation, de sélection et de filtrage, et δ_{rule} correspond à la phase d'exécution.

- La fonction de transition δ_{fact} :

  δ_{fact} (EF, IF, SF, ER, IR, SR) = (EF, IF, EF, ER+(R_E^T.EF), IR, SR)
- La fonction de transition δ_{rule} :

  δ_{rule} (EF, IF, SF, ER, IR, SR) = (EF+(RS.ER), IF, SF, ER, IR,^ER), où la matrice R_E^T désigne la transposée de R_E et ^ER désigne la négation du vecteur booléen ER.

4.3 Les étapes du processus adopté

Le processus de fouille de données adopté par notre système est composé de 6 étapes majeures :

1ere étape : Sélection et prétraitement des données

A partir des banques de données (NCBI, ...), il y'a récupération des informations biologiques relatives aux souches mentionnées ci-dessous, sous leurs formats originaux. Les agents pathogènes (souches) ciblés par cette étude sont ceux dont l'annotation a été finie à savoir : Mt H37Rv, Mt CDC1551, Mt F11, Mt H37Ra [17]. Un nettoyage, une mise en forme et une caractérisation sont effectués afin de dégager des descripteurs « attributs » possibles.

2eme étape : Transformation des données

La transformation des données du format original vers un formalisme base de données (attribut, valeur), est faite. De plus à partir des informations relatives aux RD et MIRU des séquences en question, il est défini d'autres caractéristiques calculables ou non, s'en suivra alors une «binarisation».

3eme étape : Production et évaluation des règles d'associations

La recherche des Items, des Itemsets et des règles d'association, est faite par l'algorithme Apriori [22] avec calcul systématique du support et de la confiance pour chaque règle pour ne retenir que celles ayant le support et la confiance dépassant les valeurs fixées par l'utilisateur.

4eme étape : Transformation

Les règles trouvées sont transformées puis représentées selon un formalisme transitoire aidant à la production d'un graphe d'induction. Ainsi la règle d'association R_i se verra traduite en une règle booléenne transitoire selon le principe suivant :

(R_i, Antécédent, Conséquent, support, confiance, note)

↓ ↓ ↘

(R_i , $Prémisse_i$ (Antécédent) , $Conclusion_i$ (Conséquent))

5eme étape : Production du graphe d'induction

Un graphe d'induction est construit selon le principe suivant : Un sommet désigne un nœud sur lequel on fait un test, avec les résultats possibles binaires ou à valeur multiple.

6eme étape : Représentation Cellulaire

❶ Génération des règles cellulaires à partir du graphe d'induction sous la forme :

Ri : Si Prémisse (Antécédent) alors Conclusion (Conséquent)

où Prémisse est composée des items (Itemset) de l'Antécédent de la règle d'association et la conclusion est composée des items (Itemset) de Conséquent de la règle d'association.

❷ Représentation cellulaire : Les règles générées auparavant (6.1) sont représentées en couches cellulaires. Schématiquement nous aurons :

{Ri} → REGLES et {$Prémisse_i$, $Conclusion_i$ } → FAITS

7eme étape : Intégration

Ainsi, la machine cellulaire intégrera et exploitera la représentation cellulaire et les matrices d'E/S à travers une inférence en chaînage avant pour enrichir la base de connaissances.

La dynamique de la machine cellulaire utilise les deux fonctions de transition citées auparavant (4.2).

5 Exemple d'illustration de l'induction des règles booléennes inductives.

Le processus général que notre système d'apprentissage applique à un échantillon est illustré par un exemple à partir de la 3ème étape. Nous supposons avoir obtenu les 4 règles d'association suivantes, avec les gènes (aceA, pstS, rpsG, aroK,…..etc) :

3eme étape : Production des règles d'associations

(R1,{aceA-2=1}, {pstS-3=0}, 45%,77%)

(R2,{aceA-2=0}, {rpsG=1, aroK=1}, 80%,95%)

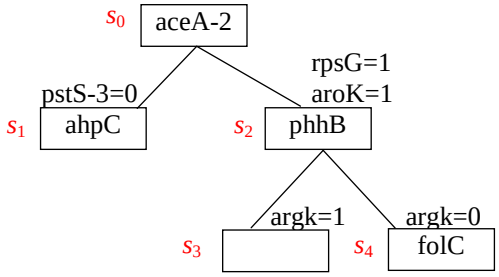
(R3,{aceA-2=0, phhB=1}, {argK=1}, 80%,70%)

(R4,{aceA-2=0, phhB=0}, {argK=0}, 45%,77%)

4eme étape : Transformation

R1,{aceA-2=1}, {pstS-3=0}
R2,{aceA-2=0}, {rpsG=1, aroK=1}
R3,{aceA-2=0, phhB=1}, {argK=1}
R4,{aceA-2=0, phhB=0}, {argK=0}

5eme étape : Production du graphe d'induction



6eme étape : Représentation Cellulaire

❶ Génération des règles cellulaires

R1: Si $\{s_0\}$ Alors {pstS-3=0, s_1}
R2: Si $\{s_0\}$ Alors {rpsG=1, aroK=1, s_2}
R3: Si $\{s_2\}$ Alors {argK=1, s_3}
R4: Si $\{s_2\}$ Alors {argK=0, s_4}

❷ Représentation des règles cellulaires

Les couches CELFACT et CELRULE.

CELFACT
s_0
pstS-3=0
s_1
rpsG=1
aroK=1
s_2
argK=1
s_3
argK=0
s_4

FAITS		
EF	IF	SF
1	0	0
0	1	0
0	0	0
0	1	0
0	1	0
0	0	0
0	1	0
0	0	0
0	1	0
0	0	0

CELRULE
R1
R2
R3
R4

REGLES		
ER	IR	SR
0	1	1
0	1	1
0	1	1
0	1	1

Notons pour CELFACT :
EF(i)=1 : un fait déjà établi,
EF(i)=0 : un fait à établir,
IF(i)=1 : un fait du type attribut=valeur,
IF(i)=0 : un fait du type sommet.

Initialement toutes les cellules de CELFACT sont à l'état EF=0 (passif) sauf EF(1)=1, c'est la base de faits initiale.

Pour CELRULE :
Toute cellule de CELRULE est considérée règle candidate, c'est-à-dire participe à l'inférence si sa valeur=1, sinon, si sa valeur=0.

Les matrices d'E/S

R_E	R1	R2	R3	R4
s_0	1	1		
pstS-3=0				
s_1				
rpsG=1				
aroK=1				
s_2			1	1
argK=1				
s_3				
argK=0				
s_4				

R_S	R1	R2	R3	R4
s_0				
pstS-3=0	1			
s_1	1			
rpsG=1		1		
aroK=1		1		
s_2		1		
argK=1			1	
s_3			1	
argK=0				1
s_4				1

Pour les matrices d'E/S :
La matrice d'entrée R_E : si le fait i∈ à Prémisse de Rj alors $R_E(i,j) = 1$
La matrice de sortie R_S : si le fait i∈ à Conclusion de Rj alors $R_S(i,j) = 1$

6 Implémentation

Le schéma illustré par la figure 3, montre le système en termes de fonctionnalités sans pour autant fixer une quelconque chronologie pour les opérations.

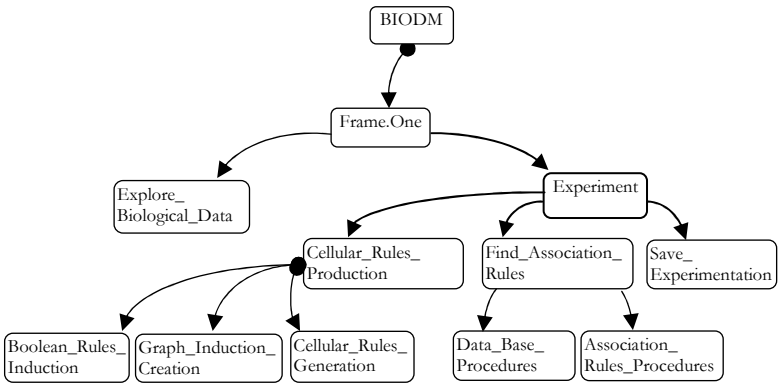


Fig. 3 : Architecture du système

1. Classe BIODM

C'est la classe qui lance toute l'application. Elle ne contient qu'une instance de la classe FRAME_ONE.

2. Classe FRAME_ONE

C'est la fenêtre principale de l'application. C'est la classe la plus importante car elle gère toutes les opérations que l'on peut effectuer.

3. Classe EXPLORE_BIOLOGICAL_DATA

Visualise les données expérimentales pour une possible vérification visuelle avant de lancer l'expérimentation.

4. Classe EXPERIMENT

Démarre l'expérimentation en demandant à l'utilisateur de sélectionner les fichiers nécessaires. Elle fait appel à des méthodes stockées telles que LECTURE_FICHIER_SEQUENCE, et CALCUL_FREQUENCE, pour le calcul des évaluations des règles.

5. Classe FIND_ASSOCIATION_RULES

Recherche les règles d'associations. Elle fait appel à des méthodes stockées dans la classe DATA_BASE_PROCEDURES, et la classe ASSOCIATION_RULES _PROCEDURES. Elle présente les résultats sous la forme textuelle, et permet de sauvegarder l'expérimentation par le biais de la classe SAVE_EXPERIMENT.

6. Classe SAVE_EXPERIMENT

Sauvegarde les résultats de l'expérimentation.

7. Classe DATA_BASE_PROCEDURES

Regroupe toutes les méthodes de gestion de la base de données, telles que la création d'une connexion, l'écriture dans une table, les requêtes sur les différentes tables, etc.

8. Classe ASSOCIATION_RULES _PROCEDURES

Recherche les itemsets, calcule les supports et les fréquences, et produit les règles d'association.

9. Classe CELLULAR_RULE_PRODUCTION

Produit les règles cellulaires et utilisant au besoin les méthodes de stockage des classes GRAPHE_INDUCTION_CREATION et CELLULAR_RULES_GENERATION.

10. Classe BOOLEAN_RULES_INDUCTION

Produit les règles booléennes inductives, en utilisant au préalable des méthodes pour la transformation des règles d'association trouvées.

11. Classe GRAPHE_INDUCTION_CREATION

Crée le graphe d'induction à l'aide d'un algorithme approprié et des règles d'associations produites à l'étape 9. Ce graphe sera le paramètre d'entrée de la classe CELLULAR_RULES_GENERATION.

12. Classe CELLULAR_RULES_GENERATION

Produit les règles cellulaires et les intègre dans la base de connaissances de la machine cellulaire CASI. Cette classe regroupe toutes les méthodes de création des couches CELFACT et CELRULE.

7 Expérimentation

En se basant sur les données expérimentales des souches citées auparavant (4.3 1^{iere} étape), les différentes expériences nous donnent des résultats intéressants qui resterons à consolider avec de nouvelles souches en cours de séquençage et qui seront prises en considération par notre système au fur et mesure de leur publication définitive sur leurs sites d'origines (NCBI, …). Pour le reste, c'est à dire la représentation booléenne (CASI), elle va certainement changer ou plus tôt « évoluer » en fonction des résultats (i.e. règles d'association produites).

8 Conclusion

Notre étude se voulait être assez novatrice, dans la mesure où nous avons été incités à utiliser des techniques prouvées de la machine cellulaire CASI [1], [3], [4], combinées à une fouille de données. De ce fait, deux objectifs nous ont guidés dans la proposition d'un automate cellulaire pour l'optimisation, la génération, la représentation et l'utilisation d'une base de règles d'association. En effet, le premier c'est d'avoir une base de règles optimisée et des temps de traitements assez réduits grâce aux principes de représentation cellulaire, et le deuxième c'est d'apporter une contribution à la construction des systèmes à base de connaissances en adoptant une nouvelle technique cellulaire. Ainsi, les avantages de notre méthode basée sur la machine cellulaire CASI peuvent être récapitulés comme suit :

- Un prétraitement simple et minimal de la base de règles d'association, pour sa transformation en matrice binaire selon le principe de couches cellulaires.
- La facilité d'implémentation des fonctions de transitions δ_{fact} et δ_{rule} qui sont de basses complexités, efficaces et robustes et concernent des valeurs extrêmes, et bien adaptées aux situations avec beaucoup d'attributs.
- La possibilité de description de l'état initial mais aussi de classifier en vu de produire des résultats simples à être insérés et utilisés à nouveau par un système expert classique, grâce notamment au système de prédiction de CASI, composé d'un ensemble de fonctions de transitions et de règles de production simples, et aussi à une facilité de transformation et de simplification des règles à travers la matrice d'incidence R_E.

Références

[1] Atmani, B., Beldjilali, B. (2007). Knowledge Discovery in Database : Induction Graph and Cellular Automaton. Computing and Informatics Journal, Vol. 26 N°2 171-197.

[2] Abbello, J., Cormode, G. (2006). Mining and Epidemiolgy (DIMACS Workshops).

[3] Abdelouhab, F., Atmani, B. (2008). Intégration automatique des données semi-structurées dans un entrepôt cellulaire, Troisième atelier sur les systèmes décisionnels, 10 et 11 octobre 2008, Mohammadia – Maroc, pp. 109-120.

[4] Benamina, B., Atmani, B. (2008). WCSS: un système cellulaire d'extraction et de gestion des connaissances, Troisième atelier sur les systèmes décisionnels, 10 et 11 octobre 2008, Mohammadia – Maroc, pp. 223-234.

[5] Carbonnelle, B., Dailloux, M., Lebrun, L., Maugein, J.,Pernot, C. (2003). Cahier de formation en biologie médicale N°29.

[6] Chen, H., Fuller, S.S., Friedman, C., Hersh, W. (2003). Knowledge management, data mining, and text mining in medical informatics (Medical Informatics, volume 8, Springer US).

[7] Chervitz, S.A., Hester, E.T., Ball, C., Dolinski, K., Dwight, S.S., Haris, M.A., Juvik, G., Malekian, A., Roberts, S., Roe T., Scafe, C., Shroeder, M., Sherlock, G., Weng, S., Zhu, Y., Cherry, J.M., Botstein, D. (1999). Using the Sacharomyces genome databases (SGD) for

analysis of protein similarities and structure (Nucleic Acids Research, Vol 27 N° 1).
[8] Fleiishman, R.D, Alland, D., Eisen, J.A, Carpenter, L., White, O., Petersen J., Deboy, R.,Dodson, R. Gwinn M., Haft, D., Hickey, E., Kolonay, J.F., Nelson, W.C., Umayam, L.A., Ermolayeva, M., Salzberg, S.L., Delcher, A., Utterback, T., Weidman, J., Khouri, H., Gill, J., Mikula, A., Bishai, W., Jacobs, W.R., Venter, J.C., Fraser, C.M. (2008). Whole-Genome comparaison of Mycobacterium Tuberculosis clinical and laboratory stains. (BMC Medical Genomics).
[9] Ferdinand, S., Valetudi, G., Sola, C., Rastogi, N. (2004). Data mining of Mycobacterium Tuberculosis complexe genotyping results using mycobacterial intersepted repetitive units validates the clonal structure of spolygotyping-defined families.
[10] Hergalant, S., Aigle, B., Leblond, P., Mari, J.F., Decaris, B. (2002). Fouille de données à l'aide de HMM : application à la détection de réitérations intragénomiques (jobim).
[11] Hergalant, S., Aigle, B., Leblond, P., Mari, J.F. (2005). Fouille de données du génome à l'aide de modèles de Markov Cachées (EGC).
[12] Hergalant, S., Eng, C., Thibessar, A., Leblond, P, Mari, J. (2005). Data mining using Hidden Markov Models (HMM2) to detect heterogeneicities into bacterial genome. (Jobim).
[13] Loria équipe Orpailleur Inserm U525 Equipe 4. (2005). Combinaison de méthodes symboliques-numériques de fouilles de données pour l'étude et l'analyse de la cohorte Stanislas (jobim).
[14] Labie, D. (2003). Le génome des mycobactéries : étude biologique et interprétation évolutive (M/S n° 3, vol. 19).
[15] Mhamdi, F., Elloumi, M., Rakotomalala, R. (2006). Extraction et sélection des n-grammes pour le classement des protéines (EGC).
[16] Maumus, S., Napoli, A., Szathmary, L., Visvikis-Siest, S. (2005). Fouille de données biomédicales complexes : Extraction de règles et de profils génétiques dans le cadre de l'étude du syndrome métabolique (Jobim).
[17] National Center for Biotechnolgy Information. : http://www.ncbi.nlm.nih.gov
[18] Niyaz, A., Hasnain, S.E. (2004). Genomics of Mycobacterium Tuberculosis: Old threats newtrends (Indian Journal Med Res 120, pp 207-212).
[19] Organisation Mondiale de la santé. : http://www.who.int/fr/
[20] Yokoyama, E., Kishida, K., Ishinohe, S. (2007). Improved Molecular Epidemiological nalysis of Mycobacterium Tuberculosis Strains Using Multi-Locus Variable Number of Tandem Repeats typing(Jpn. J. Infect. 60).
[21] Zaki, MJ., Wang, J.T.L., Toivonen, H.T.T. (2002). Recent Advances in Data Mining for Bioinformatics (BIOKDD).
[22] Zucker, J.D. (2008). Introduction à la fouille de données en bioinformatique (cours master EID-P13. IRD UR GEODES).

Abstract. The advent of new biotechnologies has led, in recent years, accumulating data on the genomes of pathogens epidemiology. As against the exploitation of genomic data do not follow the pace of discovery, then the search of biological data, particularly epidemiological nature has imposed itself to help find some answers to questions arises that the epidemiologist on specific diseases.

Hence, the problem addressed by this study is that data mining of biological Mycobacterium Tuberculosis responsible for tuberculosis. We propose a process of data-enough to generate new knowledge that will be profitable and grown at two levels:

- Take advantage of the specialist field, through the extraction of particular patterns in the rules of association which help to better understand the pathology.
- Thereafter, the extracted association rules are modeled by the Boolean principle adopted by the cellular machinery CASI (Cellular Automaton for Symbolic Induction).

The purpose of this modeling by the Boolean principle to reduce the complexity of storage and response time.

Keywords: Cellular Automaton, Biological Data Mining, Rule induction, Pattern, Itemset, Association rule, Mycobacterium Tuberculosis, Tuberculosis, Epidemic, Genome, Biology.