OrderGeneMiner : Logiciel pour l'extraction et la visualisation de motifs partiellement ordonnés à partir de puces à ADN

Mickaël Fabrègue[*,****] Agnès Braud[**], Sandra Bringay[***,♯] Florence Le Ber[****], Charles Lecellier[♯♯], Pascal Poncelet[***], Maguelonne Teisseire[*,***]

[*] TETIS, Irstea, 500 Rue Jean-François Breton, 34000 Montpellier
[**] LSIIT, CNRS-Uds, Pôle API Bd Sébastien Brant, 67412 Illkirch
[***] LIRMM UM2 CNRS, UMR 5506 - CC 477, 161 rue Ada, 34095 Montpellier
[****] LHYGES ; Université de Strasbourg/ENGEES, CNRS ; 67000 Strasbourg
[♯] MIAp UM3, Université Paul-Valery, Route de Mende, 34199 Montpellier
[♯♯] IGMM, UMR5535, Montpellier

Résumé. Le Virus de l'Immunodéficience Humaine (VIH) est actuellement un problème majeur de santé publique. Depuis l'identification du VIH, plus de 20 millions de personnes ont été identifiées. Le VIH continue de ravager les populations dans le monde entier avec 3 millions de nouvelles infections par an. Contrairement au cancer, les approches de biologie intégrative sont toujours rares dans le domaine de la lutte contre le HIV. Dans cet article, nous proposons de contribuer au développement d'une telle stratégie, en présentant un logiciel de fouille de données qui va permettre d'appliquer les concepts de motifs séquentiels et de motifs partiellement ordonnés aux données de puces à ADN. Ce logiciel se focalise sur les besoins des biologistes: 1) permet à l'expert d'intéragir dans le processus d'extraction des motifs; 2) offre une visualisation des motifs extrait sous la forme d'un graphe coloré qui résume un ensemble de motifs séquentiels. Il en résulte une visualisation plus compacte et simple qui facilite l'interprétation des experts.

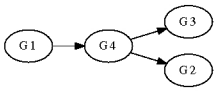
1 Introduction

L'objectif de cette étude est le développement d'un logiciel pour extraire des ensembles de biomarqueurs moléculaires (motifs de gènes) qui puissent être utilisés dans l'identification des différences entre les souches HIV1 et HIV2, qui sont deux variantes connues du virus VIH. Les biomarqueurs moléculaires sont générés à partir d'analyses de puces à ADN et sont basés sur une technique particulière de data mining : les motifs partiellement ordonnés étudiés dans Casas-Garriga et Balcázar (2004) et Acharya et al. (2007). Ils sont une extension des motifs séquentiels (motifs de gènes corrélés et ordonnés en fonction de leur niveau d'expression). Dans Salle et al. (2009), nous fournissons un algorithme pour extraire des motifs séquentiels à partir de gènes. Un exemple de motif est $\langle (G_1)(G_2, G_3) \rangle$ 80% qui signifie "Pour 80% des puces à ADN, le niveau d'expression du gène G_1 est inférieur que celui des gènes G_2 et

	4H	8H	24H	48H	72H
HIV2	$\langle(G_1)(G_4)(G_3)\rangle$	$\langle(G_1)(G_4)(G_3)\rangle$	...	...	$\langle(G_4)(G_3)(G_1)\rangle$
HIV1-R5	$\langle(G_4)(G_1)(G_3)\rangle$	$\langle(G_4)(G_3)(G_1)\rangle$	...	...	$\langle(G_3)(G_4)(G_1)\rangle$
HIV1-X4	$\langle(G_1)(G_3)(G_4)\rangle$	$\langle(G_4)(G_1)(G_3)\rangle$	...	...	$\langle(G_3)(G_1)(G_4)\rangle$

TAB. 1 – *Jeu de données VIH*

G_3, qui ont un niveau d'expression similaire. Dans Fabregue et al. (2011), nous avons montré comment ces motifs peuvent être efficacement utilisés pour la classification de tumeurs, dans le contexte du diagnostic du cancer du sein. Cependant, les motifs séquentiels sont limités car difficiles à utiliser par les experts : 1) ils sont souvent trop nombreux, 2) redondants (e.g. $\langle(G_1)(G_2)(G_3)\rangle$ et $\langle(G_1)(G_2)\rangle$ sont générés. Le second motif est inclus dans le premier et les deux motifs portent une information proche), 3) difficilement exploitables par les biologistes qui ne savent pas comment interpréter une telle liste de motifs. Dans le contexte de cet article, nous présentons une extension de ces motifs, appelés motifs partiellement ordonnés. Ceux-ci peuvent résumer un ensemble de motifs séquentiels en un seul représenté sous la forme d'un graphe. Par exemple, si les motifs $\langle(G_1)(G_4)(G_3)\rangle$ et $\langle(G_1)(G_4)(G_2)\rangle$ sont présents dans les mêmes puces à ADN, ils peuvent être résumés en un seul motif partiellement ordonné :



Avec ce logiciel, nous proposons de faciliter l'interprétation des résultats par l'expert en fournissant une visualisation sous forme de graphes. Ces derniers sont enrichis d'informations sur les gènes, comme leur fluctuation et la proximité d'expression entre eux, en jouant sur différents paramètres comme les couleurs ou bien l'épaisseur des arcs et des sommets. Ce logiciel est développé dans le cadre d'une collaboration entre le LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier) et l'IGMM (Institut de Génétique Moléculaire de Montpellier) dans le cadre d'un projet PEPS.

2 Description des données

Les données fournies sont des informations issues de l'analyse de puces à ADN où l'on a pour chaque gène la valeur de son expression. Le virus est composé de trois souches différentes qui sont HIV2, HIV1-X4 et HIV1-R5. Ces différentes souches ont été inoculées à des cellules saines clonées. Pour chacune des cellules infectées, un prélèvement a été effectué à intervalles de temps différents à 4, 8, 24, 48 et 72 heures après inoculation. Et pour chaque prélèvement, une puce à ADN a été fabriquée. Un exemple de jeu de données est représenté par la table 1 où une séquence de gènes est construite pour chaque puce à ADN (e.g. dans l'exemple trois gènes).

Nous allons maintenant présenter le protocole d'expérimentation qui va du chargement des données jusqu'à la visualisation des motifs partiellement ordonnés. Certaines étapes nécessitent l'intervention des experts pour filtrer les motifs selon leurs besoins.

	4H	8H	24H	48H	72H
HIV2	$\times$	$\times$	$\times$	$\emptyset$	$\emptyset$
HIV1-R5	$\times$	$\times$	$\emptyset$	$\emptyset$	$\emptyset$
HIV1-X4	$\times$	$\times$	$\times$	$\times$	$\times$

TAB. 2 – *Exemple de matrice d'inclusion*

	4H	8H	24H	48H	72H
HIV2	$\times$	$\times$	$\times$	$\times$	$\times$
HIV1-R5	$-$	$-$	$-$	$-$	$-$
HIV1-X4	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$

TAB. 3 – *Exemple de requête*

3 Protocole

1. Génération des séquences de données pour chaque puce à ADN. Pour générer de telles séquences, nous ordonnons les gènes en fonction de leur valeur d'expression. Ainsi, la séquence $\langle (G_3)(G_4)(G_1) \rangle$ signifie que le gène G_4 a une expression supérieure au gène G_3 mais inférieure au gène G_1.

2. Extraction des motifs séquentiels clos pour éviter la redondance de l'information. Un motif séquentiel m est clos si il n'existe pas de motif m' tel que $m \subseteq m'$ et $Frequence(m) = Frequence(m')$. Par exemple soit $m_1 = \langle (G_1)(G_2) \rangle 80\%$ et $m_2 = \langle (G_1)(G_2)(G_3) \rangle$ 80%, m_1 est inclus dans m_2 et leurs fréquences sont identiques, donc m_1 n'est pas clos.

3. Pour chaque motif extrait nous générons une matrice d'inclusion. Cette structure montre dans quelle souche et à quelle date le motif est inclus. Le tableau 2 donne un exemple d'inclusion de motif, le symbole $\times$ signifie que le motif est inclus, le symbole $\emptyset$ qu'il ne l'est pas. Ainsi dans l'exemple, le motif concerné est présent dans la souche HIV2 huit heures après inoculation du virus mais n'est pas présent dans la souche HIV1-R5 24 heures après inoculation.

4. Intervention de l'expert qui fait une requête dans l'ensemble de motifs extraits. Cette requête se fait sous la forme d'une matrice et seuls les motifs dont la matrice d'inclusion respecte la requête sont conservés. Le tableau 3 est un exemple de requête, le symbole $\times$ signifie que les motifs filtrés doivent être présents dans la souche et l'heure concernée, le symbole $\emptyset$ signifie qu'ils ne doivent pas l'être et le symbole $-$ lorsque les motifs peuvent être présents ou non. Ainsi la requête en exemple sert à filtrer les motifs qui sont présent à toutes les dates de la souche HIV2 mais dans aucune de la souche HIV1-X4.

5. Génération d'un motif partiellement ordonné à partir des motifs filtrés par l'expert. Pour générer un tel motif, nous utilisons la bibliothèque graphique Graphviz de Ellson et al. (2001). Elle offre la possibilité de générer facilement des graphes en énumérant la liste des sommets et des arcs qui les relient. En plus du graphe, nous ajoutons des informations sur les gènes :

Couleur des sommets Un sommet de couleur bleue signifie que le gène a vu sa valeur d'expression augmenter au cours des différents temps d'inoculation. Une couleur rouge décrit une valeur d'expression qui a au contraire diminuée.

Épaisseur des sommets Plus le sommet est épais, plus la valeur d'expression à variée au cours des différents temps d'inoculation (variation positive ou négative).

Épaisseur des arcs entre les sommets Plus les gènes sont éloignés par leur valeur d'expression moyenne, plus l'arc sera épais.

La taille des graphes générés est égale au nombre total de gènes présents dans les motifs séquentiels filtrés. Nous donnons l'exemple d'un tel graphe en annexe. Il a été généré à partir de la requête qui consiste à ne filtrer que les motifs présents dans la souche HIV2 mais non présent dans la souche HIV1-X4, peut importe HIV1-R5 (requête de l'exemple 3).

4 Conclusion

Dans ce travail, nous proposons un logiciel pour la visualisation efficace de motifs séquentiels sous forme de graphes dans le contexte des puces à ADN. Nous permettons aux biologistes d'intéragir dans le processus de fouille de données en filtrant les motifs désirés. Enfin l'application affiche les résultats sous forme d'un graphe enrichi qui résume les motifs séquentiels extraits pour chaque requête. Nous avons pour perspective d'utiliser le logiciel sur d'autres jeux de données biologiques comme le cancer, en collaboration avec l'IRCM (Institut de Recherche en Cancérologie de Montpellier).

Références

Acharya, M., T. Xie, J. Pei, et J. Xu (2007). Mining api patterns as partial orders from source code : from usage scenarios to specifications. In *Proceedings of the the 6th joint meeting of the European software engineering conference*, ESEC-FSE '07, New York, NY, USA, pp. 25–34. ACM.

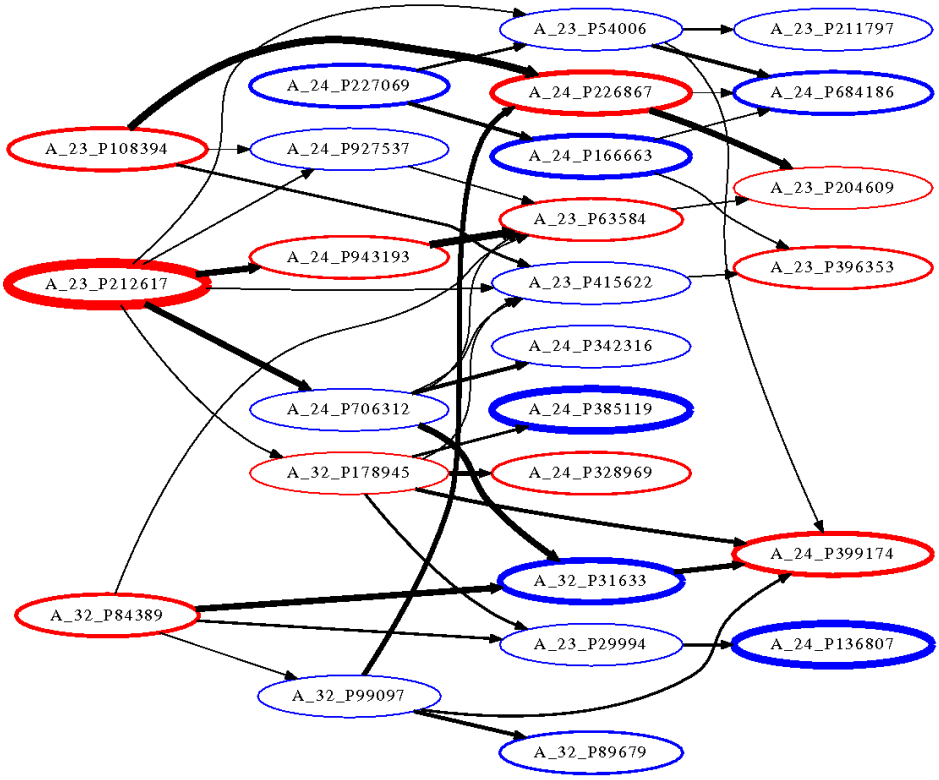
Casas-Garriga, G. et J. L. Balcázar (2004). Coproduct transformations on lattices of closed partial orders. In *ICGT*, pp. 336–351.

Ellson, J., E. Gansner, L. Koutsofios, S. North, G. Woodhull, S. Description, et L. Technologies (2001). Graphviz - open source graph drawing tools. In *Lecture Notes in Computer Science*, pp. 483–484. Springer-Verlag.

Fabregue, M., S. Bringay, P. Poncelet, M. Teisseire, et B. Orsetti (2011). Mining microarray data to predict the histological grade of a Breast Cancer. *Journal of Biomedical Informatics*.

Salle, P., S. Bringay, et M. Teisseire (2009). Demon : Decouverte de motifs séquentiels pour les puces adn. In *EGC*, pp. 459–460.

Annexe



HIV2 vs X4

Summary

HIV (Human immunodeficiency virus infection) is a major public health issue today. Since the identification of HIV, more than 20 million people have been identified. HIV continues to ravage populations around the world with 3 million of new infections per year. Unlike cancer, Integrative Biology approaches are still rare in the field of HIV. In this paper, we propose to contribute to the development of these strategies by presenting a data mining software which apply sequential patterns and partial order patterns on DNA microarrays data. This software focus on biologists needs: 1) allows experts to interact in the pattern extraction process; 2) provides a visualization of extracted patterns in a colored graph that summarizes a sequential pattern set. This results in a more compact and simple visualization that facilitates experts' interpretation.