[PGROU - Iguana]
Outil pour la visualisation de sous-composantes d'un graphe, à l'aide de la programmation logique, destiné au pronostic des patients atteints du cancer.

Keywords Info : Intelligence Artificielle, Data Mining, Programmation Logique, Machine Learning, Graphes
Keywords Developement : Python, CyRest, Cytoscape, R

Contexte

Les techniques de fouille de données (ou *Data Mining*) consistent à extraire les caractéristiques les plus représentatives d'un grand jeu de données. Dans ce contexte ils existent, parmi les plus connus, les méthodes de clustering et, parmi ceux dont nous sommes le plus familiarisés dans notre recherche en informatique, des méthodes sur les concepts formels [1]. L'outil **Iguana** (Fig. 1), développé récemment dans le cadre d'un projet PAPPL, permet de visualiser et rendre accessible aux utilisateurs non-experts un analyse implémenté par une méthode IGGY-POC basé dans la programmation logique [2] qui utilise des concepts formels pour construire *k sous-graphes* (Fig 2) à partir d'un graphe principal. Ces k sous-graphes ont la particularité d'encapsuler un type d'information logique qui nous intéresse (coloration du graphe avec 2 couleurs en respectant une propriété entre la cible et ses prédécesseurs), réduisant efficacement l'exploration exhaustive de toutes les coloration possibles du graphe. IGGY-POC trouve ces composants utilisant des contraintes qui optimisent la vérification d'une propriété locale du graphe écrites sous forme de programme logique et résolues ensuite à l'aide des solveurs pour l'Answer Set Programming [3]. Nous avons évalué comment IGGY-POC permet d'extraire de composants du graphe issues de la base de données PID-NCI [3] (environ 30000 arcs et 18000 noeuds) qui compile ~300 réseaux de signalisation[1] chez l'humain. Grace à un jeu de données du cancer Myélome Multiple (MM) produit par le laboratoire CRCINA[2] de 600 patients, nous avons observé que certains composants de ce graphe permettaient de distinguer et mieux caractériser ces données pour comparer les types cellulaires atteints du MM vs normal. Nous avons mise en valeur ces résultats par rapport à d'autres méthodes du clustering connus.

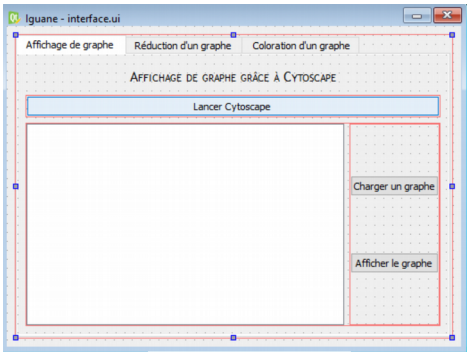

Figure 1: Iguana

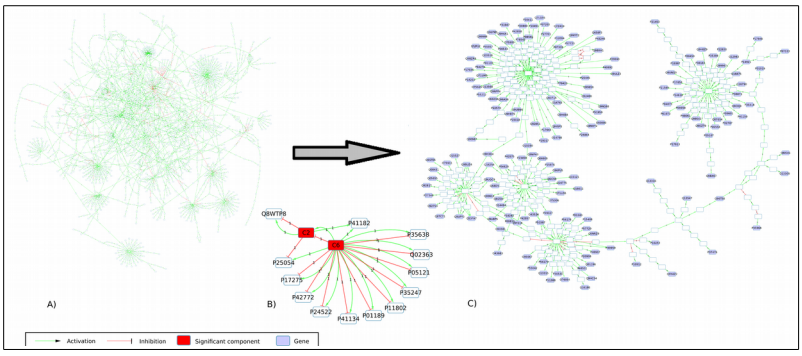

Figure 2: IGGY-POC, réduction du graphe A) en 15 composantes illustrées en B) dont une illustrée en C)

1 Un graphe qui décrit les chaines biochimiques existants à l'intérieur d'une cellule, à travers l'activation de différents protéines qui déclenchent la transcription de divers gènes lorsque les récepteurs de cette cellule sont stimulés par des facteurs extérieurs ou stress, comme des nutriments, des rayons UV, etc.
2 Centre de Recherche en Cancérologie et Immunologie Nantes Angers

Objectifs du travail

1. Permettre à Iguana de charger une matrice M de donnés transcriptomiques[3]. Les données que nous utiliserons dans ce projet seront issus du DREAM Challenge[4], qui est un regroupement à niveau international destiné à mettre à disponibilité des données cliniques et des mesures expérimentales à la communauté méthodologique pour évaluer leur méthodes et comparer leur résultats. Les données que nous étudierons dans ce projet correspondent à la mesure de l'expression de gènes pour plusieurs patients atteints du MM et pour lesquelles l'information bon et mauvaise pronostic est connue.
2. Inclure en Iguana la fonctionnalité du calcul de la similarité entre la coloration des k composantes-extraites et les états des noeuds du graphe définis par une colonne de matrice M.
3. Inclure dans Iguana la possibilité de classifier les profils d'expression de la matrice M en construisant des modèles bon vs. mauvais pronostic à l'aide des techniques du Random Forest, à partir de la matrice de similarité précédemment implémenté dans l'étape 2. Un pipeline a été déjà développé et mis à disposition à travers de Docker[5].
4. Inclure dans Iguana la possibilité de tester la précision de ces classificateurs avec de jeux de données de test du DREAM pour des patients avec bon et mauvais pronostic.
5. Améliorer les paramètres du classificateur, ainsi que les étapes pour générer ce classificateur, pour améliorer la précision.
6. Optionnellement, développer un exécutable de Iguana multi-plateforme (MacOS et Linux).

Environnement de travail (outils à maitriser avant ou pendant le projet)

- Python, CyRest, Docker, R

Groupe de travail

- Encadrants:
    - Carito Guziolowski (MCF ECN, Equipe ComBi, Laboratoire de Sciences du Numérique de Nantes)
    - Dr Bertrand Miannay

References

[1] Poelmans, J., Ignatov, D. I., Kuznetsov, S. O., and Dedene, G. (2013). Formal concept analysis in knowledge processing : A survey on applications. Expert Systems with Applications, 40(16) :6538–6560.

[2] Miannay B, Minvielle S, Roux O, Magrangeas F, Guziolowski C: Constraints On Signaling Networks Logic Reveal Functional Subgraphs On Multiple Myeloma OMIC Data. Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2017, Boston, MA, USA, August 20-23 2017, pp 768-69

[3] Gebser, M., Kaminski, R., Kaufmann, B., Schaub, T.: Answer Set Solving in Practice. Synthesis Lectures on Artificial Intelligence and Machine Learning 6(3), 1–238 (2012)

[4] Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., Buetow, K.H.: PID: the Pathway Interaction Database. Nucleic acids research 37(Database issue), 674–9 (2009).

3 Une matrice m x n avec des valeurs réels compris entre 0 et 1 qui represente l'expression de n gènes pour m patients atteints du MM

4 https://www.synapse.org/#!Synapse:syn6187098

5 https://www.docker.com/