Atelier Fouille de Grands Graphes : Application à la bioinformatique

Organisateurs : Mohamed Elati (ISSB - Université d'Evry),
Blaise Hanczar (LIPADE - Université Paris Descartes),
Lydia Boudjeloud-Assala (LITA EA 3097 - Université de Lorraine)

PRÉFACE

Le groupe de travail Fouille de Grands Graphes a été créé en 2010, il s'intéresse à l'analyse et de l'étude de la dynamique dans les grands graphes. L'objectif du groupe est de proposer une structure d'animation scientifique pour des chercheurs venant de plusieurs disciplines et s'intéressant à la fouille de grands graphes. La recherche en modélisation, en analyse et en fouille de grands graphes a connu un net regain d'intérêt qui peut se justifier par les deux points suivants : dans plusieurs domaines les données se présentent souvent sous forme structurée : graphes ou objets reliés. Nous pouvons citer par exemple les systèmes biologiques, le web, les réseaux sociaux (facebook, twitter, ...) ou encore les réseaux bibliographiques. Les graphes issus de ces domaines ont des propriétés spécifiques qui les différencient des graphes aléatoires (graphes sans échelle, faible densité, faible degré de séparation, ...). Les techniques actuelles permettent d'observer de très grands réseaux qui évoluent dans le temps et nous posent ainsi le défi du passage à l'échelle et la nécessité de disposer d'outils de visualisation et de fouille pouvant s'adapter à ce nouveau type de données.

Pour sa troisième édition, dans le cadre de la conférence Extraction et Gestion de Connaissances (EGC'2014), le groupe de travail EGC Fouille de Grands Graphes (EGC-FGG) propose un atelier AFGG-EGC'2015 avec comme thème principal "*bioinformatique*".

L'édition 2015 de l'atelier a pour but de réunir les chercheurs intéressés par le traitement, la fouille et la visualisation de grands graphes et leurs application dans les problématiques en bioinformatique. Notre ambition est de permettre aux participants d'aborder les problèmes rencontrés liés aux données biologiques et de présenter leurs approches de traitement.

L'atelier concerne aussi bien les notions conceptuelles, théoriques que les applications abordées notamment dans une présentation invitée. Cet atelier s'intéressera aussi aux réseaux biologiques, aux graphes métaboliques avec une application autour des réseaux d'interaction protéiques.

Nous tenons à remercier les auteurs, les membres du comité de programme pour la qualité de leurs contributions ainsi que les responsables des ateliers d'EGC 2015.

Enfin nous remercions vivement les présidents : Jérôme Darmont, président du comité de programme, Benoît Otjacques et Thomas Tamisier co-présidents du comité d'organisation d'EGC 2015.

Mohamed Elati	Blaise Hanczar	Lydia Boudjeloud-Assala
Université d'Evry,	Université Paris Descartes,	Université de Lorraine,
ISSB	LIPADE	LITA EA 3097

Membres du comité de lecture

Le Comité de Lecture est constitué de:

Lydia Boudjeloud
Bruno Cremilleux
Mohamed Elati
Blaise Hanczar
Philippe Leray
Amedeo Napoli
Celine Rouveirol
Nataliya Sokolovska

TABLE DES MATIÈRES

Enumération et statistiques pour l'étude des réseaux biologiques

Etienne Birmelé *

* Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes
etienne.birmele@parisdescartes.fr,
http://www.math-info.univ-paris5.fr/ ebirmele/

1 Résumé

Les réseaux biologiques sont des objets très divers et de plusieurs ordres de grandeur plus petits que d'autres réseaux comme ceux issus d'internet ou des réseaux sociaux. Ils présentent cependant la difficulté d'avoir des arêtes qui ne peuvent être directement observées mais doivent être inférées à base d'informations hétérogènes. De plus, les mécanimes biologiques qu'ils résument sont pour la plupart encore assez mal connus.

L'étude de tels réseaux présente par conséquent de nombreux problèmes statistiques, comme leur inférence ou l'établissement de modèles aléatoires de référence pouvant servir de base de comparaison pour des études de sur-représentation. De plus, de par leur complexité et l'incertitude qui leur est inhérente, résoudre un problème sur ces réseaux en se contentant de trouver la solution maximisant une fonction objectif pourrait mener à un contre-sens biologique. Il est souvent préférable de lister plusieurs solutions intéressantes, avec pour objectif de les valider/infirmer à posteriori sur la base de critères biologiques, ce qui donne lieu à des problématiques liées à l'énumération de structures.

Le but de cet exposé est de préciser ces différentes problématiques et de donner des exemples à la fois de solutions et de questions ouvertes qui y sont liées.

2 Présentation de l'orateur

Après une thèse en théorie des graphes sous la direction d'Adrian Bondy, Etienne Birmelé a été Maître de Conférences au laboratoire *Statistique et Génome* d'Evry. Il s'est tourné vers l'analyse des réseaux biologiques, à la fois d'un point de vue statistique en travaillant notamment avec Christophe Ambroise, et d'un point de vue plus algorithmique à travers une collaboration soutenue avec l'équipe de Marie-France Sagot à l'INRIA Rhône-Alpes.

Il est depuis l'automne 2013 Professeur au sein du laboratoire de Mathématiques Appliquées de Paris Descartes (MAP5) et responsable du Master d'Ingénierie Mathématique pour les Sciences du Vivant au sein de cette université.

Ses principaux centres d'intérêts sont le dévelopement de modèles de graphes aléatoires et leur apprentissage d'un point de vue statistique, ainsi que l'étude des algorithmes d'énumération de structures d'intérêt dans les réseaux biologiques.

Summary

The formalized study of biological networks implies several statistical problems, as their inference or the choice of a random model. Morevoer, it is biologically more relevant to enumerate all maximal solutions of optimisation problems rather than only finding the maximum one, leading to enumeration problems. This talk will present different examples of solutions and open problems about those questions.

Graphe métabolique pour la recherche de chemins conservés de transformations chimiques

Maria Sorokina*,**,*** David Vallenet*,**,***

*Commissariat à l'Energie Atomique et aux Energies Alternatives, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, 91057, Evry, France
**CNRS-UMR8030, 2 rue Gaston Crémieux, 91057, Evry, France
***UEVE, Université d'Evry Val d'Essonne, bd François Mitterrand, 91057, Evry, France
msorokina@genoscope.cns.fr

Résumé. Dans cet article, nous proposons une nouvelle représentation d'un réseau métabolique sous la forme d'un graphe orienté de transformations chimiques. Partant d'un réseau de réactions, les noeuds du graphe de transformation regroupent des réactions partageant une même signature moléculaire et les arcs sont établis à partir de la connectivité initiale des réactions. Différents scores sont ensuite calculés pour évaluer la conservation de chemins métaboliques déjà connus et de découvrir de nouveaux chemins.

1 Introduction

Dans le domaine de la bioinformatique, le métabolisme est généralement modélisé sous forme de graphes orientés dont les noeuds représentent des réactions et/ou des métabolites et les arêtes des échanges de produits/substrats entre les réactions (Lacroix et al., 2008). Pour un organisme donné, la reconstruction du réseau métabolique se fait généralement à partir de l'annotation de son génome qui prédit des activités enzymatiques et donc la présence de réactions. Cette reconstruction par homologie est limitée par les difficultés d'assigner des fonctions correctes aux gènes à partir de leur séquence ADN et également par l'absence de caractérisations expérimentales d'activités enzymatiques dont les enzymes sont pour certaines inconnues (Sorokina et al., 2014). Des sous-ensembles de ces graphes sont souvent utilisés pour représenter des voies métaboliques qui regroupent un ensemble de réactions participant à un même processus biologique d'importance pour l'organisme. Il existe plusieurs grandes hypothèses sur l'origine et l'évolution des voies métaboliques, dont l'hypothèse de l'évolution en patchwork par recrutement des enzymes dans les nouvelles voies métaboliques (Jensen, 1976; Ycas, 1974), la synthèse rétrograde qui postule que la construction des voies métaboliques s'effectue à partir du métabolite final (Horowitz, 1945) ou encore la duplication des voies métaboliques (Huynen et al., 2000; Schmidt et al., 2003). Malgré leurs différences, elles se rejoignent sur un point qui est la capacité des enzymes de catalyser un ou plusieurs types de réaction sur des substrats plus ou moins différents, phénomène aussi appelé promiscuité enzymatique. Une étude récente (Notebaart et al., 2014) a mis en évidence cette capacité des enzymes à s'adapter à de nouveaux substrats chez *Escherichia coli*.

L'objectif du travail présenté dans cet article est d'explorer le métabolisme en recherchant des enchaînements de transformations chimiques qui sont supposés conservés au cours de l'évolution. La détection de ces unités chimiques ou modules réactionnels permettra, ainsi, de prédire de nouvelles voies métaboliques dont les réactions et les enzymes ne sont pas ou partiellement connues. Une méthode (Muto et al., 2013), ayant un but similaire, a été publiée en 2013 et a permis d'identifier un certain nombre de ces modules conservés en réalisant un alignement de voies métaboliques et en se basant sur un regroupement des réactions réalisant les mêmes motifs de transformations chimiques, appelés RClass (Kotera et al., 2004). Ici, nous adoptons un formalisme différent pour la recherche de conservation d'enchainements de transformations chimiques. L'exploration se fait sur un réseau métabolique réduit à des signatures de transformations chimiques à la place des réactions. Ce formalisme en graphe nous permet ainsi de définir plusieurs métriques de conservation suivant le nombre de voies métaboliques concernées, le nombre de réactions, le nombre d'enzymes associées ou l'importance topologique des noeuds et des arêtes. Des scores utilisant ces métriques sont ensuite calculés pour des chemins métaboliques connus et pour tous les chemins possibles du réseau.

2 Matériel et méthodes

2.1 Réseau de réactions

Les données utilisées ici sont extraites de la base de données MetaCyc (Caspi et al., 2014) qui contient des voies métaboliques curées de tous les domaines du vivant. En plus des voies métaboliques y sont répertoriés les métabolites, les réactions, les enzymes et les gènes liés. Les voies métaboliques décrites dans MetaCyc sont assez courtes (en moyenne 4 à 5 réactions) et ont été expérimentalement élucidées. A partir de ces voies métaboliques, un réseau de réactions a été construit. Dans ce réseau orienté, un noeud représente une réaction et deux noeuds sont reliés par un arc si le produit de la première réaction est le substrat de la deuxième. Seules les réactions appartenant à une voie métabolique ont été prises en compte, car seulement dans ces dernières les composés chimiques sont classés comme primaires ou secondaires, permettant de faire la distinction entre les métabolites principaux de la réaction et les co-substrats d'aide à la réaction tels que l'eau, l'ATP, NADP, etc. La connexion est donc basée uniquement sur les composés chimiques primaires ce qui permet d'éviter un réseau excessivement connexe et de ne créer des liens qu'entre les réactions en s'appuyant exclusivement sur les métabolites d'intérêt.

2.2 Signatures moléculaires de réactions

La méthode des signatures moléculaires de réactions utilisée (RMS) (Carbonell et Faulon, 2010) permet de calculer le type de transformation chimique à partir de la différence de structure entre les produits et les substrats impliqués dans la réaction. Les réactions effectuant le même type de transformation chimique auront la même RMS. A partir d'une RMS, il est possible de prédire toutes les réactions possibles qui réalisent ce type de transformation chimique.

2.3 Signatures moléculaires de réactions et les familles de protéines

Afin de calculer la conservation des RMS d'un point de vue conservation des protéines au travers de tous les organismes ayant leur génome complètement séquencé, nous avons étudié les correspondances entre les familles de protéines Pfam (Finn et al., 2014) et les RMS. Deux sources d'information ont été utilisées : les numéros EC (Bairoch, 1994), reliant les protéines de la banque UniProt (The UniProt Consortium, 2014) aux réactions, et les annotations de MetaCyc. Trois métriques basées sur le nombre de protéines reliées aux RMS par leurs familles Pfam sont calculées. Les ratios *p2r* et *r2p* représentent la fraction de protéines appartenant à une famille et associées à une RMS donnée parmi respectivement toutes les protéines de la famille (*p2r*) et toutes les protéines associées à la RMS (*r2p*). Le score d'association d'une famille de protéines à une RMS (*score(pf,rms)*) est calculé à l'aide de la moyenne harmonique entre *p2r* et *r2p*. Ce score évalue la sensibilité et la spécificité qu'une protéine, appartenant à une famille Pfam donnée, catalyse un type de réaction métabolique décrit par une RMS donnée et, réciproquement, qu'une protéine effectuant un type de réaction appartienne à une famille de protéines donnée. La valeur de cette métrique tend à être très basse lorsque les effectifs de la famille de protéines et/ou de la RMS sont importants.

2.4 Réseau de transformations chimiques

Le réseau de réactions est transformé en un réseau orienté de transformations chimiques représentées par les RMS (*FIG 1*). Les noeuds dans le réseau de réactions sont regroupés selon les RMS. Deux RMS sont reliées dans ce réseau si il existe au moins un arc reliant une réaction ayant pour signature la première RMS et une réaction ayant pour signature la deuxième RMS. Si la RMS source est identique à la RMS d'arrivée l'arc n'est pas créé.

2.5 Scores de conservation de chemins de transformations chimiques

Dans le réseau de RMS, plusieurs poids ont été placés sur les noeuds et les arcs. Sur les noeuds, il y a le nombre de réactions MetaCyc que regroupe la RMS en question, qu'elles soient présentes dans le réseau initial de réactions ou non (*poidsRea*). On associe aussi aux noeuds un poids (*poidsProt*) représentant la diversité des protéines effectuant un type de réaction donné du point de vue de leur conservation inter organismes : la moyenne, pour chaque famille Pfam associée à une RMS, du nombre de protéines de la famille, issues de protéomes complets, multiplié par le *score(pf,rms)* correspondant. Sur les arcs est placé un poids topologique (*poidsTopo*) basé sur la moyenne harmonique du score PageRank (Brin et Page, 1998) et du ratio de liens entre les réactions que les liens entre RMS représentent (*FIG 1*). Ce poids permet de retrouver les noeuds, les arcs et les chemins intéressants du point de vue topologie dans le réseau de RMS, notamment les chemins à forte influence. Une énumération de tous les chemins de longueurs deux à quatre dans le réseau de RMS a été effectuée en utilisant la librairie java Grph (Hogie, 2013). Pour que les résultats puissent être comparables, les voies métaboliques issues de MetaCyc et traduites en RMS on été découpées en chemins chevauchants de longueurs deux à quatre. Un score pour les chemins (métaboliques et issus d'énumération) est calculé, pour chaque poids définis précédemment, en faisant une moyenne géométrique entre les poids topologiques (*scoreTopo*) et une moyenne arithmétique pour les poids de conservation de protéines (*scoreProt*) et les poids de conservation de réactions (*scoreRea*).

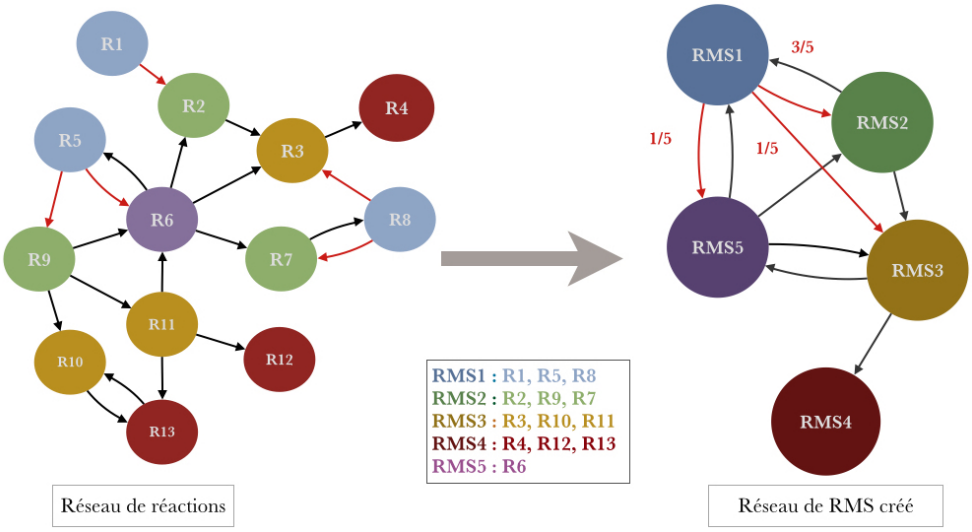

FIG. 1 – *Exemple de transformation d'un réseau de réactions en un réseau de transformations chimiques (RMS). Les réactions appartenant à une même RMS sont représentées par des noeuds d'une même couleur. Le score, indiqué en rouge sur les arcs, indique le ratio de liens entre les réactions que représentent les liens entre RMS.*

3 Résultats et discussion

3.1 Réseau de réactions versus réseau de transformations chimiques

Dans le réseau de réactions, on compte 7538 noeuds et 17624 arcs ; le degré moyen d'un noeud est de 4,48. Dans le réseau de transformations chimiques, il y a 4957 noeuds regroupant 8722 réactions, dont 1184 non incluses dans le réseau initial de réactions, et 10436 arcs ; le degré moyen d'un noeud est de 4,03. Il y a environ 1,2 réactions par RMS, en sachant que plus de la moitié des RMS ne sont signature que d'une seule réaction. Dans le réseau de RMS, il existe 49291 chemins de longueur 2 (deux arcs et trois noeuds), 289159 chemins de longueur 3 et 2195101 chemins de longueur 4.

3.2 Chemins conservés de transformations chimiques

La conservation d'un chemin dans le réseau de RMS a été évaluée par le calcul de trois scores à partir de poids sur les noeuds ou les arcs du graphe (*scoreRea, scoreProt, scoreTopo*). Ces scores ont été déterminés pour tous les chemins de longueurs 2 à 4. Les chemins inclus dans les voies métaboliques connues de MetaCyc ont été identifiés. Pour ces derniers, un score supplémentaire de conservation a été calculé et représente le nombre de voies métaboliques connues de MetaCyc dans lesquelles on retrouve le chemin de RMS (*scoreMeta*).

Parmi les voies métaboliques connues, le *scoreRea* maximal est 43,6, avec une moyenne de 1,40. Ce même score, parmi tous les chemins possibles, a pour maximum 47 et une moyenne

de 1,24. Le maximum parmi tous les chemins pour le *scoreProt* s'élève à 8559,4 (moyenne 700,8), et pour les voies métaboliques à 6796,2 (moyenne 793,6). Le *scoreTopo* maximal est de 0,00179 (moyenne globale à 0,00066), mais parmi les voies métaboliques connues son maximum n'est que de 0,00134, avec une moyenne de 0,00050. D'après ces chiffres, on peut dire qu'il existe des chemins métaboliques fortement conservés qui ne sont pas encore représentés dans les bases de données comme MetaCyc. On fera ici la remarque que la notion de voie métabolique n'est qu'une représentation humaine et subjective des enchainements potentiels de réactions dans une cellule vivante, permettant de découper, et non pas de partitionner, le métabolisme en blocs plus faciles à étudier. On considère que plus le score d'un chemin est élevé, plus il est conservé. En fixant la p-valeur à un seuil de 5% pour les scores de chemins de longueur 2, le seuil est de 1809,1 pour le *scoreProt* et de 0,0009384 pour le *scoreTopo*. Pour le *scoreRea*, on considère qu'il y a conservation quand ce score est supérieur ou égal à 2, c'est à dire que les RMS du chemin considéré sont signatures d'au moins deux réactions en moyenne. Après avoir découpé les voies métaboliques en chemins chevauchants de longueur 2 pour que les scores soient comparables, on a déterminé que parmi les 928 voies métaboliques de plus de trois réactions, 154 sont conservées selon le *scoreRea*, 87 selon le *scoreProt* et 46 selon le scoreTopo. Il y a cependant très peu d'intersections entre les chemins conservés selon ces scores : seulement 12 sont conservés à la fois selon les scores *scoreRea* et *scoreProt*, 18 selon *scoreRea* et *scoreTopo*, 2 selon *scoreProt* et *scoreTopo*, et seulement une voie métabolique est conservée d'après les trois scores. Il s'agit du "superpathway of purine nucleotide salvage".

Il y a 491 chemins de longueur 2 dans le réseau de RMS qui sont retrouvés dans au moins deux voies métaboliques distinctes (*scoreMeta* supérieur ou égal à 2). En retour, 402 voies métaboliques (parmi les 928 exploitables pour ces calculs) issues de MetaCyc contiennent au moins un de ces chemins, donc plus de 40% des voies sont au moins en partie constituées de blocs conservés de transformations chimiques.

4 Conclusion

Dans cet article nous proposons une nouvelle méthode permettant d'explorer le métabolisme et de détecter des chemins de transformations chimiques que nous supposons conservés au cours de l'évolution. Nous avons pu déterminer les voies métaboliques conservées selon différentes métriques, ce qui permettra par la suite d'en découvrir des nouvelles en explorant les chemins les plus conservés parmi tous les chemins possibles dans le réseau construit.

Références

Bairoch, A. (1994). The ENZYME data bank. *Nucleic acids research 22*(17), 3626–3627.

Brin, S. et L. Page (1998). The Anatomy of a Search Engine.

Carbonell, P. et J. L. Faulon (2010). Molecular signatures-based prediction of enzyme promiscuity. *Bioinformatics 26*(16), 2012–2019.

Caspi, R., T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang, et P. D. Karp (2014).

The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research 42*(D1).

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, et M. Punta (2014). Pfam : The protein families database.

Hogie, L. (2013). Grph :The high performance graph library for Java.

Horowitz, N. H. (1945). On the Evolution of Biochemical Syntheses. *Proceedings of the National Academy of Sciences of the United States of America 31*(6), 153–157.

Huynen, M., B. Snel, W. Lathe, et P. Bork (2000). Exploitation of gene context.

Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annual review of microbiology 30*, 409–425.

Kotera, M., Y. Okuno, M. Hattori, S. Goto, et M. Kanehisa (2004). Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *Journal of the American Chemical Society 126*(50), 16487–16498.

Lacroix, V., L. Cottret, P. Thébault, et M.-F. Sagot (2008). An introduction to metabolic networks and their structural analysis. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM 5*(4), 594–617.

Muto, A., M. Kotera, T. Tokimatsu, Z. Nakagawa, S. Goto, et M. Kanehisa (2013). Modular architecture of metabolic pathways revealed by conserved sequences of reactions. *Journal of Chemical Information and Modeling 53*(3), 613–622.

Notebaart, R. A., B. Szappanos, B. Kintses, F. Pal, A. Gyorkei, B. Bogos, V. Lazar, R. Spohn, B. Csorg, A. Wagner, E. Ruppin, C. Pal, et B. Papp (2014). Network-level architecture and the evolutionary potential of underground metabolism. *Proceedings of the National Academy of Sciences 111*(32), 11762–7.

Schmidt, S., S. Sunyaev, P. Bork, et T. Dandekar (2003). Metabolites : a helping hand for pathway evolution ? *Trends in biochemical sciences 28*(6), 336–41.

Sorokina, M., M. Stam, C. Médigue, O. Lespinet, et D. Vallenet (2014). Profiling the orphan enzymes. *Biology direct 9*, 10.

The UniProt Consortium (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic acids research 42*(Database issue), D191–8.

Ycas, M. (1974). On earlier states of the biochemical system. *Journal of theoretical biology 44*(1), 145–160.

Summary

In this paper we propose a new representation for a metabolic network as a network of chemical transformations. Starting from a reaction network, reactions having the same molecular signature are grouped in nodes and edges are established from the initial reaction connectivity. Different scores are then computed in order to evaluate the conservation of known metabolic pathways and to discover new pathways.

Détection de zones denses en triplets significatifs dans un réseau orienté

Nicolas Dugué*, Tennessy Kolubako*, Anthony Perez*

*Univ. Orléans, INSA Centre Val de Loire, LIFO EA 4022, FR-45067
prenom.nom@univ-orleans.fr

Résumé. L'étude des triplets dans un réseau orienté donne des informations significatives sur celui-ci. En effet, chaque famille de réseaux -du web, biologiques- possède un profil particulier. Ce profil est caractérisé par la présence en plus grand ou plus faible nombre de certains types de triplets par rapport à des réseaux aléatoires possédant la même distribution de degrés. Nous nous intéressons à la fouille de zones denses en triplets significatifs. Pour cela, nous rappelons la notion de profil de siginificativité et l'appliquons à une large variété de réseaux du réel et artificiels. Pour détecter des zones denses en triplets, nous proposons une approche basée sur la détection des coeurs de communautés du graphe de co-occurence des triplets, issu du réseau initial.

1 Introduction

L'étude des voisinages d'un noeud (*e.g.* coefficient de clustering) dans un réseau complexe permet d'obtenir des informations pertinentes sur sa topologie locale, mais des motifs plus complexes fournissent une information plus riche. Dans cette optique, Milo et al. [8] étudient la topologie des réseaux complexes en s'intéressant notamment aux triplets formés par les liens orientés du réseau (Figure 1) .

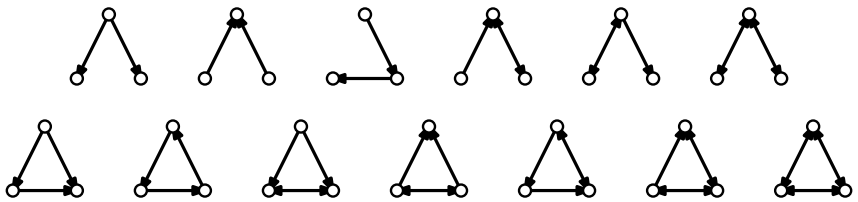

FIG. 1 – *Les 13 triplets orientés possibles.*

Ces motifs constituent des blocs caractéristiques d'un type de réseau et de la façon dont il se construit. Cette notion est généralisée par Milo et. al [7] qui parlent alors de *familles de réseaux*. Chaque famille de réseaux est caractérisée par un profil de significativité statistique. Ce profil est établi par comparaison, pour chaque type de triplet, du nombre de triplets détectés

dans le réseau étudié au nombre de triplets obtenu en moyenne dans des graphes aléatoires possédant une même distribution des degrés. Plus précisément, pour chaque type de triplet, Milo et. al [7] calculent la significativité statistique en utilisant d'abord la notion de Z-score :

$$Z_i = \frac{r_i - \mu(a_i)}{\sigma(a_i)}, \text{ avec } i \in \{1, .., 13\} \quad (1)$$

où i représente le type de triplet étudié, r_i le nombre de triplets de type i détectés sur le réseau étudié, $\mu(a_i)$ le nombre de triplets de type i détectés en moyenne sur un ensemble de réseaux aléatoires et $\sigma(a_i)$ l'écart-type à cette moyenne pour le même ensemble de réseaux aléatoires. Le profil de significativité statistique du réseau est ensuite calculé en normalisant le vecteur des Z-score obtenu pour chaque type de triplet :

$$PS_i = \frac{Z_i}{\sqrt{\sum Z_i^2}}, \text{ avec } i \in \{1, .., 13\} \quad (2)$$

Ainsi, des vecteurs PS fortement corrélés indiquent que les réseaux associés font partie de la même famille. Milo et. al [7] constatent que réseaux sociaux et réseaux du web forment une seule famille. Ils mettent en évidence deux familles supplémentaires : celle des réseaux biologiques issus de microorganismes ou encore celle des réseaux d'adjacence de mots.

Contributions Dans un premier temps, nous appliquons la méthode développée par Milo et. al [7] à un grand nombre de réseaux orientés de différents types et tailles (notamment issus du LFR benchmark [4]). Par ailleurs, nous posons les bases d'une méthode pour détecter les zones denses en motifs significatifs dans un réseau. Cette méthode est basée sur la détection des coeurs de communautés du graphe de co-occurence des noeuds du réseau pour un type de triplet. Nous nous intéressons également à la densité des sous-graphes détectés par la méthode des coeurs, une question qui n'avait pas été étudiée à notre connaissance.

2 Des familles de réseaux

Afin d'établir le profil de significativité statistique de réseaux du réel et issus du LFR benchmark [4], nous détectons les triplets de chaque type en utilisant l'algorithme de Latapy [5]. Nous générons ensuite un ensemble de graphes aléatoires conservant la distribution des degrés du réseau étudié et calculons le profil selon l'équation (2).

2.1 Les réseaux du réel

La Figure 2 présente les corrélations entre les profils de significativité statistique de 25 réseaux complexes issus de la base de données KONECT [1]. Nous considérons des réseaux :

- d'adjacence de mots (Spanish_book, French_book, Darwin_book)
- sociaux dont les liens indiquent la méfiance ou la confiance (Epinions et Slashdot)
- sociaux du web, notamment les liens entre utilisateurs sur Twitter, Flickr, Livejournal et Youtube
- de référence entre pages web (le site de l'Université Notre-Dame, le moteur de recherche Baidu et l'encyclopédie chinoise Hudong)

— de communications par mail (Enron, University of California Irvine, une manufacture, ou au sein d'une institution de l'UE)
— physiques comme ceux reliant des aéroports (US airports et Openflight)
— biologiques issus de [7], ou encore de co-occurence d'achats Amazon, de réponses sur des forums Slashdot.

Les résultats observés confirment les résultats obtenus par Milo et. al [7]. Les réseaux du web récents (et donc plus conséquents que ceux étudiés dans [7]), tels que ceux issus de conversation via des threads ou des forums, certains réseaux de communications par email et les réseaux sociaux du web possèdent un profil de significativité statistique proche de celui des réseaux sociaux à taille humaine et des réseaux de pages web, identifié par Milo et. al [7]. Cependant, certains réseaux de communication par e-mail possèdent une structure proche des réseaux d'adjacence de mots. Les réseaux sociaux du web où les utilisateurs désapprouvent d'autres utilisateurs suivent également ce profil, contrairement à ceux où les utilisateurs s'approuvent qui suivent eux le profil des réseaux du web. Enfin, on retrouve les réseaux biologiques ensemble.

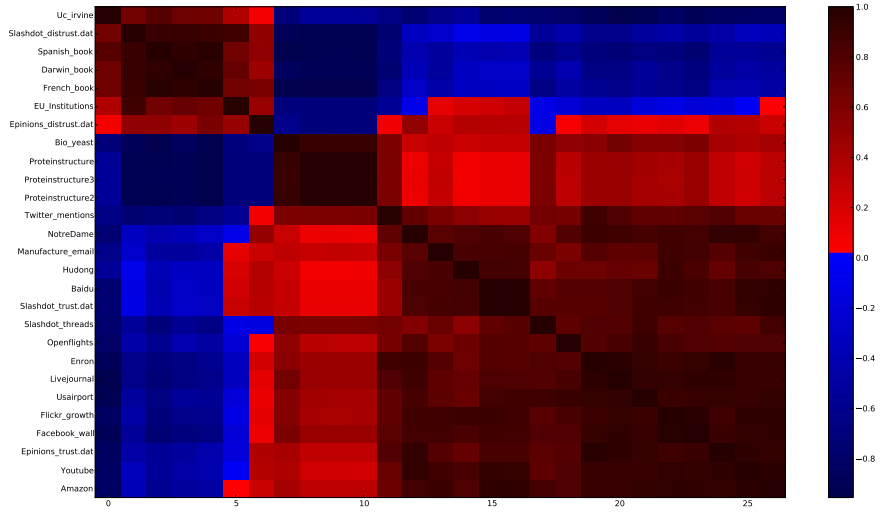

FIG. 2 – *Correlation entre différents réseaux du réel extraits de KONECT [1].*

2.2 Les réseaux artificiels

Les réseaux artificiels générés par le LFR benchmark [4] garantissent une distribution des degrés en loi de puissance, ainsi qu'une structure de communautés plus ou moins bien définie dont la taille est également distribuée selon une loi de puissance. Les paramètres utilisés dans nos expérimentations sont proches des paramètres des réseaux du réel [9]. La Figure 3 présente

les corrélations entre les profils de significativité statistique des réseaux du LFR. Le nombre de sommets est fixé à 1000, et la distribution des tailles de communautés à 1. La distribution des degrés, t_2 est fixée à 2 ou 3. La taille minimale m_c (resp. maximale M_c) des communautés varie de 50 à 100 (resp. de 100 à 400). Enfin, le degré moyen k prend les valeurs 10 ou 20 quand le degré maximum M_k est de 100 ou 500. Pour chaque paramètre, 10 réseaux sont générés et nous utilisons le profil moyen. Les profils d'un réseau du web (Hudong), biologique (proteine-structure), d'adjacences de mots (French_book), de communication par mail (Uc_Irvine) sont ajoutés pour comparaison.

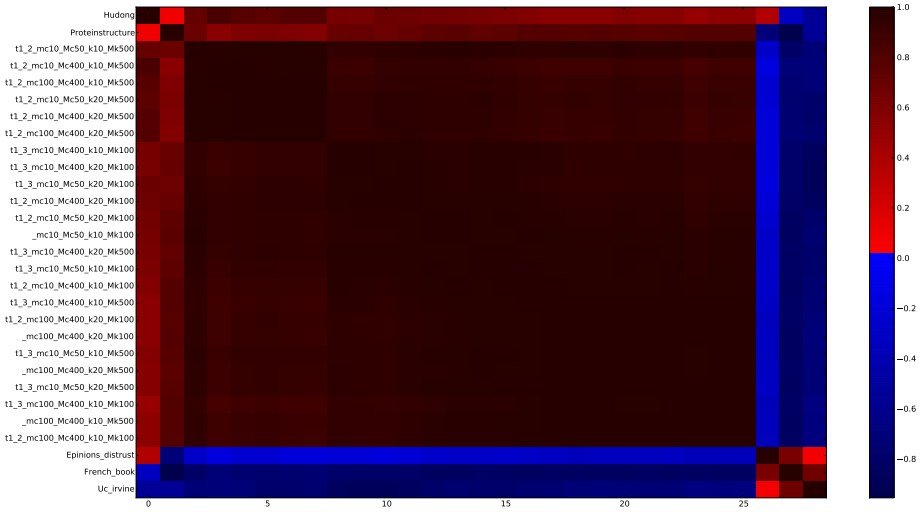

FIG. 3 – *Correlation entre les réseaux du LFR benchmark [4] et certains réseaux du réel [1].*

On observe que les réseaux générés ont des profils proches des réseaux sociaux ou du web, mais sont *anti-corrélés* avec les réseaux d'adjacence de mots, certains réseaux issus de communication par e-mail, et les réseaux tels qu'Epinions où un utilisateur peut indiquer son manque de confiance en un autre utilisateur. Puisque le réalisme des réseaux artificiels a un impact sur l'évaluation des algorithmes de détection de communautés [9], cette considération est importante.

3 A la recherche de zones denses en triplets significatifs

Le profil de significativité statistique met en évidence la plus grande importance de certains motifs dans la construction du réseau. Afin d'étudier efficacement ces motifs importants, mais sans les énumérer de façon exhaustive, nous proposons une méthode destinée à détecter les zones denses en un motif donné dans le réseau.

Coeurs de communauté Par définition, l'optimisation de la modularité [3] permet d'obtenir une partition du graphe telle que pour chaque partie, ses noeuds sont plus connectés entre eux que vers l'extérieur. Il s'agit en quelque sorte d'obtenir un ensemble de parties les plus denses possibles. Les coeurs de communautés [10] fournissent un consensus entre différentes partitions fournies par un algorithme d'optimisation non déterministe. Notre méthode, basée sur l'algorithme de détection de communautés glouton de Louvain [2] est dotée d'une bonne efficacité calculatoire, et peut donc être appliquée sur des données de taille conséquentes. Nous formalisons maintenant notre méthode.

Méthodologie Soit $D = (V, A)$ un graphe orienté. Nous nous intéressons aux triplets de D d'un type $i \in \{1, .., 13\}$. Nous commençons par détecter les triplets de ce type i en utilisant l'algorithme décrit par Latapy [5]. Un nouveau graphe valué et non-orienté $G_\Delta = (V, E_\Delta, \omega_\Delta)$ est ensuite créé, où $\omega_\Delta : E_\Delta \rightarrow \mathbb{N}$ représente les occurences des sommets de V dans les triangles détectés. Soit $(v_1, v_2) \in V \times V$, alors $(v_1, v_2) \in E_\Delta$ s'il existe au moins un triangle de type i dans G contenant v_1 et v_2. Le poids de l'arête (v_1, v_2) est alors égal au nombre de triangles dans lequel ces sommets apparaissent ensemble. Nous effectuons ensuite la détection des coeurs de communauté sur ce graphe telle que décrite dans [10]. Pour cela, nous lançons λ exécutions de l'algorithme de Louvain en mode non-déterministe sur G_Δ. Nous construisons à partir des résultats un nouveau graphe non-orienté et valué $G_C = (V, E_{Cs}, \omega_C)$ où pour tous sommets v_1, v_2 de G_Δ, l'arête (v_1, v_2) appartient à E_C s'ils font partie *au moins une fois* de la même communauté dans les différentes partitions produites par l'algorithme de Louvain. La fonction de poids $\omega_C(v_1, v_2)$ a dans ce cas pour valeur le nombre de fois où v_1 et v_2 apparaissent dans une même communauté. Nous utilisons enfin un seuil $\alpha \in [0, 1]$ pour supprimer les arêtes (v_1, v_2) de E_C vérifiant $\omega_C(v_1, v_2) < \alpha \times \lambda$, et finalement obtenir les coeurs de communautés. Ces derniers nous fournissent donc des zones denses en motifs significatifs sur le graphe initial.

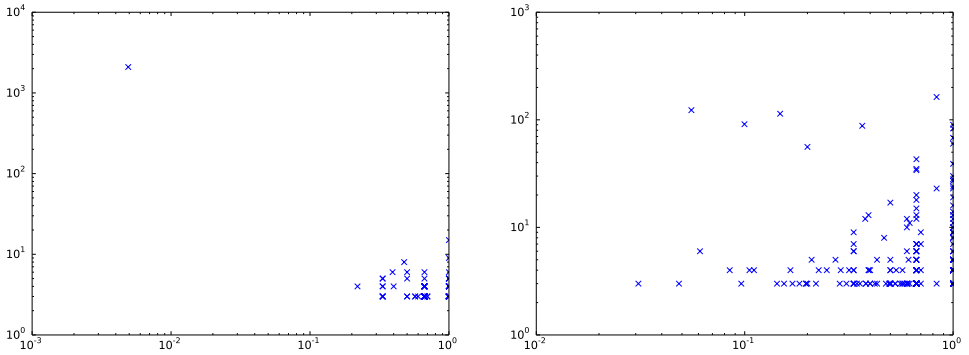

FIG. 4 – *Distribution de la densité des sous-graphes détectés (abscisse) par la méthode des coeurs de communautés en fonction de leur taille en nombre de sommets (ordonnée) sur le réseau Petster [1]. A gauche,* $\alpha = 0.1$*, à droite,* $\alpha = 0.9$*.*

Comme le montre la Figure 3, elle permet de paramétrer la recherche afin d'obtenir des zones plus ou moins denses, ou de taille plus ou moins grande. Enfin, contrairement aux algo-

rithmes de détection de quasi-cliques [6], la méthode tient compte des poids des arêtes dans le recherche de sous-graphes denses.

4 Conclusion

Dans cet article, nous étudions tout d'abord les profils de significativité statistique d'un grand nombre de réseaux du réel de taille plus conséquente que ceux étudiés par Milo et. al [7], et de catégories différentes. Nos expérimentations montrent en particulier que les réseaux artificiels du LFR benchmark [4] sont incapables de générer des réseaux aux profils proches de ceux des réseaux d'adjacence de mot. Ceci limite leur réalisme et peut notamment impacter l'évaluation de la performance des algorithmes de détection de communauté [9] sur ce type de réseaux. Par la suite, nous nous intéressons à la densité des coeurs détectés par la méthode dite des coeurs de communauté, en fonction du paramètre α utilisé. Notre objectif est de combiner ces deux observations afin de développer une méthode permettant de rechercher des sous-graphes denses en motifs significatifs.

Références

[1] KONECT datasets.

[2] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *J. of Stat. Mech.*, 2008(10) :P10008.

[3] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring User Influence in Twitter : The Million Follower Fallacy. In *ICWSM*, 2010.

[4] Andrea Lancichinetti and Santo Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Phys. Rev. E*, 80(1), 2009.

[5] Matthieu Latapy. Main-memory triangle computations for very large sparse (power-law)) graphs. *Theoretical Computer Science*, 407(1-3) :458–473, 2008.

[6] Victor E. Lee, Ning Ruan, Ruoming Jin, and Charu Aggarwal. A survey of algorithms for dense subgraph discovery.

[7] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663) :1538–1542, March 2004.

[8] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs : Simple building blocks of complex networks. *Science*, 298(5594) :824–827, 2002.

[9] G.K. Orman and V. Labatut. The effect of network realism on community detection algorithms. In *ASONAM*, pages 301–305, Aug 2010.

[10] M. Seifi. *Coeurs stables de communautés dans les graphes de terrain*. PAF, 2012.

Summary

Studying triplets in an oriented complex network can provide meaningful insight. We investigate the problem of detecting subgraphs that are dense with respect to specific kind of triplets. Our method relies on the so-called *coeurs de communautés* method applied on a specific graph derived from the original network.

PEPPER: Optimisation multi-objective pour la recherche de complexes moléculaires dans des réseaux d'interaction protéiques

C. Winterhalter*, R. Nicolle * A. Louis *, C. To * F. Radvanyi ** M. Elati *

* iSSB, University of Evry
**CNRS/Institut Curie

Résumé. Nous présentons PEPPER (Winterhalter et al., 2014), un plugin Cytoscape ayant pour objectif d'identifier des complexes protéiques en tant que sous-réseaux connectés à partir d'une graine (liste initiale) de protéines issue d'études protéomiques. PEPPER identifie des sous-graphes connectés basé sur l'optimisation de deux objectifs: (i) la couverture, une solution doit contenir autant de protéines de la graine que possible, (ii) la densité, les protéines d'une solution doivent être aussi connectées que possible, en utilisant les données connues de réseaux d'interactions à l'échelle du protéome. Les comparaisons effectuées sur des jeux de données de référence chew la levure et l'humain révèlent que l'approche intégrative de PEPPER est de qualité supérieure à celle des méthodes standards pour la découverte de complexes protéiques. La visualisation et l'interprétation des résultats est facilitée par un post-traitement automatique des solutions basé sur l'analyse topologique et l'intégration de données biologiques aux complexes protéiques. PEPPER est un outil facile d'utilisation qui peut être utilisé pour analyser n'importe quelle liste de protéines. PEPPER est disponible à partir du gestionnaire de plugins Cytoscape ou en ligne (`http://apps.cytoscape.org/apps/pepper`) sous license GNU tout public version 3.

Références

Winterhalter, C., R. Nicolle, A. Louis, C. To, F. Radvanyi, et M. Elati (2014). Pepper : cytoscape app for protein complex expansion using protein protein interaction networks. *Bioinformatics 30*, 3419–3420.

Summary

We present PEPPER (Winterhalter et al., 2014), a Cytoscape app designed to identify protein complexes as densely connected subnetworks from seed lists of proteins derived from proteomic studies. PEPPER identifies connected subgraph by using multi-objective optimisation involving two functions: (i) the coverage, a solution must contain as many proteins

from the seed as possible, (ii) the density, the proteins of a solution must contain as many interactions as possible be as connected as possible, using only interactions from a proteome-wide protein interaction network. Comparisons based on gold standard yeast and human datasets showed PEPPER's integrative approach as superior to standard protein complex discovery methods. The visualisation and interpretation of the results are facilitated by an automated post-processing pipeline based on topological analysis and data integration about the predicted complex proteins. PEPPER is a user-friendly tool that can be used to analyse any list of proteins. PEPPER is available from the Cytoscape plugin manager or `http://apps.cytoscape.org/apps/pepper` and released under GNU General Public License version 3.

Index