Proposition de thèse - 2015

Compression sans perte de graphes pour l'aide à la compréhension de réseaux biologiques

Contact et directeur de thèse: Jacques NICOLAS, DR INRIA Jacques.Nicolas[AT]irisa.fr
Equipe: Dyliss (Bioinformatics) : http://www.irisa.fr/dyliss
Financement et date de tdébut envisagés: MENRT: October 2015

Contexte: La production de données de grands graphes s'est multipliée du fait des applications internet (web, réseaux sociaux) aussi bien que des application techniques (réseaux de communication ou d'électricité) que scientifiques (physique statistique, biologie). Bien que les graphes soient typiquement dessinés en ayant recours à des techniques sophistiquées de visualisation, ces graphiques deviennent rapidement un enchevêtrement inextricable de nœuds et d'arcs lorsque la taille du graphe croît. Aider à comprendre le contenu en information des grands graphes est un challenge important qui ne peut pas simplement être résolu à l'aide d'un tracé et d'une mise en page élaborée.

Nous nous intéressons à des graphes issus de données d'observation en sciences expérimentales. Dans ce domaine, les graphes (aussi appelés réseaux) sont un moyen d'intégrer différentes sources de données et le rôle de la compression est d'aider le scientifique à extraire de la connaissance de ce type de données. Ce but général peut se décliner de nombreuses manières : simplifier la visualisation, montrer la structure globale du graphe, trouver des règles générales d'association, ou simplement rechercher une meilleure efficacité pour la réalisation d'autres tâches (par exemple la reconnaissance de motifs

Objectif: La thèse étudiera les moyens de résumer un graphe qui préservent la structure de chemins, dans le contexte des sciences expérimentales et plus particulièrement la biologie. A un haut niveau, le problème central de la compression de graphe peut être vu comme un problème de clustering. La préservation de la structure globale du graphe peut être réalisée en conservant la représentation par graphe et en remplaçant les nœuds du graphe initial par des nœuds plus abstraits correspondant à des sous-ensembles de nœuds. Deux nœuds abstraits sont reliés par un arc abstrait si tous leurs éléments sont reliés par un arc (ils forment une biclique dans le graphe initial). Les clusters de nœuds sont typiquement des nœuds partageant des arcs et des propriétés en commun [1].

La compression de graphes sans perte exige que chaque arc soit conservé et pose donc le problème de la couverture d'un graphe par un sous-ensemble de ses bicliques. L'idée la plus simple est de rechercher une partition des nœuds mais cette condition sera la plupart du temps trop exigeante et en pratique on recherche plutôt une partition des bicliques, i.e. un ensemble de sous-ensembles d'arcs mutuellement disjoints. Ce problème est connu comme étant NP-complet, même si on se restreint à des graphes bipartis, mais est soluble à paramètres fixés (FPT) [2]. L'article [3] propose un outil bioinformatique (Power Graph Analysis) qui produit par un algorithme glouton une partition de bicliques telle que les sous-ensembles de nœuds forment une hiérarchie. Il a été appliqué avec succès sur des réseaux d'interaction de protéines [4], des réseaux médicament-cible-maladie [5] et des réseaux de régulation de gènes [6]. A partir de ce point de départ, la thèse explorera un certain nombre de lignes d'amélioration :

- Dans l'ensemble des bicliques, celles qui sont maximales (aussi appelées concepts) forment un sous-ensemble intéressant qui est l'objet d'étude de l'analyse formelle de concepts (FCA). Un cadre

théorique existe pour les décrire à l'intérieur d'une structure de treillis [7]. Comment générer la couverture du graph à partir du treillis et d'opérations algébriques simples sur les concepts ?

- La plupart des méthodes utilisent une recherche gloutonne qui produit un optimum local en terme de nombre minimum de bicliques utilisées. La recherché d'un optimum global est un problème combinatoire pour lequel les méthodes à base de contraintes semblent mieux adaptées. Nous proposons d'exprimer le problème dans le cadre d'ASP (Answer Set Programming). Celui-ci offre un langage de haut niveau qui représente la spécification d'un problème à l'aide de clauses et contraintes logiques [8]. Une fois le problème correctement représenté, un solveur se charge de la résolution sans qu'il soit nécessaire d'indiquer la manière de résoudre.
- Au lieu de rechercher des similarités entre nœuds, il est possible de rechercher des similarités entre sous-graphes. Chaque nœud abstrait devient alors un sous graphe qui peut lui-même être compressé. La recherche de sous-graphes intéressants est le sujet des techniques de fouille de graphes [9] et leur intégration devrait permettre d'améliorer les taux de compression des graphes biologiques.

Environment: La thèse se déroulera dans une équipe de bioinformatique, Dyliss, concernée par les problèmes de modélisation de mécanismes moléculaires complexes dans les cellules vivantesen utilisant des formalisations qualitatives. Au travers de collaborations multiples avec des laboratoires de biologie, elle a accès à des réseaux variés (réseaux d'interaction de protéines, réseaux de régulation, réseaux de signalisation…) qui seront utilisés pour tester les concepts développés durant la thèse. Une plateforme, Genouest, donne accès à des bases de données et des logiciels bioinformatiques et offre des capacités de calcul de type cluster ou cloud. Ce travail inclut une collaboration avec une équipe d'informatique à Potsdam qui a mis au point un des meilleurs solveurs ASP à ce jour http://potassco.sourceforge.net/ .

Mots clés: Graphes, Compression, Problèmes NP-complets, Recherche combinatoire, Optimization discrète, Analyse Formelle de Concepts, Programmation par ensembles réponse (ASP), Bioinformatique

First references:

[1] Zhang Y, Phillips CA, Rogers et al. 2014. On finding bicliques in bipartite graphs: a novel algorithm and its application to the integration of diverse biological data types. BMC Bioinformatics. 15:110. [2] Fleischner H, Mujuni E, Paulusma D, Szeider S. 2009. Covering graphs with few complete bipartite subgraphs. Theoretical Computer Science, Vol. 410 n°21-23, pp 2045-2053.
[3] Royer L, Reimann M, Andreopoulos B, Schroeder M (2008) Unraveling Protein Networks with Power Graph Analysis. PLoS Comput Biol 4(7): e1000108. doi:10.1371/journal.pcbi.1000108
[4] Royer L, Reimann M, Stewart AF, Schroeder M. Network compression as a quality measure for protein interaction networks. PLoS One. 2012;7:e35729.
[5] Daminelli S1, Haupt VJ, Reimann M, Schroeder Drug repositioning through incomplete bi-cliques in an integrated drug-target-disease network. Integr Biol (Camb). 2012 Jul;4(7):778-88.
[6] Taylor-Teeples M, Lin L, de Lucas M, Turco G et al., An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature. 2015 Jan 29;517(7536):571-5.
[7] M.Ganter B, Wille R. 1997. Applied lattice theory: Formal concept analysis. In General Lattice Theory, G. Grätzer editor, Birkhäuser. http://www.math.tu-dresden.de/~ganter/psfiles/concept.ps
[8] Lifschitz V. 2008. What is answer set programming?. In *Proceedings of the 23rd nat. conf. on Artificial Intelligence* (AAAI'08), Anthony Cohn (Ed.), Vol. 3. AAAI Press 1594-1597.
[9] S. Parthasarathy, S. Tatikonda, D. Ucar A Survey of Graph Mining Techniques for Biological Datasets. Managing and Mining Graph Data Advances in Database Systems Vol 40, 2010, pp 547-580