Math. & Sci. hum. / Mathematical Social Sciences (−−e année, n° –, —-(-), p. 1–3)

RECHERCHE DE CLASSES EMPIÉTANTES DANS UN GRAPHE : APPLICATION AUX RÉSEAUX D'INTERACTIONS ENTRE PROTÉINES

Lucile DENŒUD-BELGACEM [1]

RÉSUMÉ – *Cet article présente une méthode de classification empiétante permettant de mettre en évidence des zones denses en arêtes dans un graphe. On cherche plus précisément à extraire du graphe des sous-graphes dont la densité en arêtes soit élevée par rapport à la densité du graphe entier, ces sous-graphes pouvant avoir des sommets en commun. Cette méthode est appliquée à un problème issu de la biologique : l'annotation des protéines. Les graphes considérés traduisent alors des interactions observées entre les protéines. Partant du principe biologique que des protéines impliquées dans une même fonction cellulaire interagissent, les sous-graphes obtenus par l'application de la méthode de classification empiétante aux réseaux d'interactions donnent des indications sur les fonctions des protéines constituant ces sous-graphes, ce qui permet de fournir une aide informatique à la prédiction de fonctions inconnues de certaines protéines. Le caractère empitétant autorisé par la méthode présentée ici permet en particulier de prendre en compte le fait que les protéines peuvent être impliquées chacune dans plusieurs fonctions cellulaires.*

MOTS-CLÉS – Graphe, Partitionnement, Nuées dynamiques, Classification par densité, Classification empiétante, Réseau d'interactions entre protéines

SUMMARY – Computation of overlapping classes in a graph : application to protein-protein interactions networks

This article describes a method of overlapping classification, in order to compute zones which are dense in edges in a graph. More precisely, the aim is to compute subgraphs whose densities in edges are large with respect to the edge-density of the whole graph, while these subgraphs may share common vertices. This method is applied to a problem arising in biology : the annotation of proteins. The graphs then represent the interactions observed between proteins. Thanks to the biological principle that proteins involved in a same cellular function interact, the subgraphs provided by the application of the method to the protein-protein interactions networks give information on the functions of proteins belonging to these subgraphs. This provides a computer-aided tool for the prediction of unknown functions of some proteins. The overlapping allowed by the method depicted here permits to take into account the fact that each protein may be involved into several cellular functions.

KEYWORDS – Graph, Partitioning, k-means, Classification by density, Overlapping classification, Protein-protein interactions networks

[1] FuturMaster, 696 rue Yves Kermen, 92100 Boulogne-Billancourt, lucile.belgacem@futurmaster.com

1. INTRODUCTION

Dans cet article, nous proposons une méthode permettant de mettre en évidence des zones denses en arêtes dans un graphe. Il s'agit en fait de déterminer une classification empiétante des sommets du graphe de façon que les sous-graphes induits par ces classes soient relativement denses en arêtes par rapport au graphe global. Cette étude est motivée par une problématique biologique : l'annotation des protéines.

Dans cette introduction, nous commençerons donc par situer ce travail dans le contexte biologique, en présentant brièvement le problème de l'annotation des protéines, puis les travaux antérieurs s'inscrivant dans une démarche similaire à la nôtre. Enfin nous décrirons plus précisément l'objectif et les enjeux de la méthode proposée.

La partie 2 sera consacrée ...

Les composants de cette méthode sont ensuite ...

Cette application à des données réelles fait l'objet de la partie 4. On y analyse les propriétés des classes obtenues d'abord d'un point de vue combinatoire (partie 4.1) puis d'un point de vue biologique (partie 4.2).

Une courte conclusion (partie 5) résume les principaux éléments de l'article.

1.1. Contexte biologique[2].

Après le séquençage de la majorité des génomes des organismes utilisés en laboratoire, les biologistes sont face à un nouveau défi : comprendre, à grande échelle, la fonction des composants codés par le génome et notamment des protéines. Sachant qu'aucune protéine n'assure seule sa fonction et que la vie de chaque organisme dépend de dizaines de milliers d'interactions entre protéines spécifiques, il est nécessaire de décrire en premier lieu les interactions moléculaires entre les protéines afin d'appréhender les mécanismes complexes survenant au sein de la cellule et de l'organisme.

and so on...

2. CONCLUSION

La méthode de classification présentée ici est issue d'une problématique biologique : étant donné un réseau d'interactions entre protéines, former des classes de protéines interagissant fortement entre elles, et correspondant donc à des complexes de protéines associés à des fonctions cellulaires, l'objectif final étant de prédire les fonctions inconnues de certaines protéines. Les classes doivent pouvoir être chevauchantes car une protéine est susceptible d'intervenir dans plusieurs fonctions cellulaires. La méthode proposée ne repose pas sur l'optimisation d'une fonction objectif, trop délicate à mettre au point, et le nombre de classes n'est pas connu à l'avance. On souhaite simplement que la méthode détecte des classes intrinsèques au graphe de départ, quitte à ne pas classer tous les sommets, et que les classes obtenues aient des caractéristiques pertinentes d'un point de vue biologique (cardinal pas trop important, empiètement modéré) afin de faciliter leur interprétation.

[2]Voir par exemple [1] ou [2] pour des ouvrages génériques du domaine.

La méthode proposée se déroule en trois étapes. Tout d'abord on crée des noyaux initiaux des classes en sélectionnant les optima locaux d'une fonction de densité locale définie sur les sommets du graphe. On améliore ensuite ces noyaux par une adaptation de la méthode des nuées dynamiques qui modifie les noyaux ainsi que leur nombre. Ces deux premières étapes forment des classes disjointes. Dans la troisième et dernière étape, les noyaux sont étendus de manière empiétante suivant un critère sur la qualité des classes obtenues.

...

Références

[1] B. Alberts, D. Bray, J. Lewis, N. Carlier, C. Butor, A. Kahn (1995) *Biologie moléculaire de la cellule*, Flammarion.

[2] J. Etienne, F. Millot (1998) *Biochimie génétique, biologie moléculaire*, abrégé de médecine, Masson, 1998.