Classes empiétantes dans un graphe et application aux interactions entre protéines

Lucile Denoeud, Irène Charon, Alain Guénoche, Olivier Hudry

UNIVERSITÉ PARIS 1 – PANTHÉON - SORBONNE –

CERMSEM

UMR CNRS 8095

Cahiers de la MSE

Classes empiétantes dans un graphe et application aux interactions entre protéines

Lucile DENOEUD, CERMSEM & ENST

Irène CHARON, ENST

Alain GUENOCHE, Institut de Mathématiques

Olivier HUDRY, CERMSEM & ENST

2005.32

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Maison des Sciences Économiques, 106-112 boulevard de L'Hôpital, 75647 Paris Cedex 13

http://mse.univ-paris1.fr/Publicat.htm

ISSN : 1624-0340

Classes empiétantes dans un graphe et application aux interactions entre protéines

L. Denoeud[1,2], I. Charon[1], A. Guénoche[3], and O. Hudry[1,2]

[1] École nationale supérieure des télécommunications,
46, rue Barrault, 75634 Paris cedex 13
denoeud@enst.fr, charon@enst.fr, hudry@enst.fr
[2] CERMSEM CNRS-UMR 8095, MSE, Université Paris 1 Panthéon-Sorbonne,
106-112, boulevard de l'Hôpital, 75647 Paris cedex 13
[3] Institut de Mathématiques de Luminy, 163, avenue de Luminy, 13009 Marseille
guenoche@iml.univ-mrs.fr

Résumé Dans cet article, on étudie une méthode de classification reposant sur une recherche de zones denses en arêtes dans un graphe non pondéré. On ne cherche pas à faire un partitionnement mais à extraire des classes intrinsèques aux données, qui pourront donc être empiétantes. La méthode proposée est ensuite appliquée à un graphe d'interactions entre protéines, les classes mises en évidence pouvant permettre aux biologistes de prédire les fonctions cellulaires de certaines protéines.

Mots-Clefs. Bio-informatique ; classification ; graphe d'interactions.

1 Introduction

L'objectif de la classification est de structurer ou organiser un ensemble d'objets (*taxonomical units*) en formant des classes homogènes et bien séparées ([1], [3], [12]). L'homogénéité veut dire que des objets appartenant à une même classe se ressemblent fortement, la séparation signifie que deux objets appartenant à deux classes différentes sont dissemblables. Malheureusement, cet objectif est difficile à atteindre, car les problèmes posés en terme d'optimisation d'un critère sont en général NP-difficiles (voir par exemple [3] ou [10] ; pour la théorie de la complexité, voir [11] ou [2]) et les solutions optimales de ce type de problème restent inaccessibles en pratique. Diverses démarches ont donc été développées pour se rapprocher le plus possible de ces optima ([12]).

Dans cet article, on étudie une méthode de classification peu contraignante où on ne cherche pas forcément à classer tous les objets, ni même à former des classes disjointes, le but étant avant tout de mettre en évidence des classes intrinsèques aux données. Le point de départ est un graphe $G = (X, E)$ où les sommets sont les objets à classer et les arêtes (non pondérées) correspondent à une ressemblance entre leurs extrémités. On cherche alors des zones denses en arêtes dans ce graphe correspondant à des ensembles d'objets se ressemblant fortement.

Nous reprenons la méthode de recherche de classes disjointes dans un graphe proposée par Colombo *et al* dans [7] et [8]. Elle se déroule en deux étapes : tout d'abord on crée des noyaux à partir d'une fonction de densité définie sur les sommets du graphe, ensuite on étend les noyaux en respectant certains critères sur la qualité des classes obtenues. Nous étendons cette méthode en permettant à un sommet d'appartenir à plusieurs classes lors de la seconde étape.

Notre étude s'inscrit dans un projet interdisciplinaire sur les interactions protéine-protéine (IMPBio, projet EIDIPP) ; on s'intéresse donc à une application de la méthode à un graphe d'interactions entre protéines chez la levure. Notre objectif est de fournir des outils mathématiques et informatiques aux biologistes afin qu'ils puissent prédire les fonctions cellulaires de certaines protéines qui interagissent pour former des complexes auxquels correspondent des fonctions. Dans ce cadre il est pertinent de construire des classes empiétantes étant donné que les protéines interviennent dans plusieurs fonctions cellulaires ([**?**]).

Dans la partie 2, on décrit l'algorithme et on propose de nouveaux critères d'extension des noyaux. Dans la partie 3, on applique l'algorithme à un graphe d'interactions entre protéines, et on compare les différents critères d'extension. On validera la méthode sur une application. Enfin on conclut sur la méthode proposée.

2 Algorithme de classification par densité

Les méthodes de classification par densité ont été introduites par Wishart en 1976 ; il s'agit de construire les classes autour d'éléments qui possèdent beaucoup de proches voisins ([13], [6]). Le but de l'algorithme est de détecter des zones denses en arêtes dans un graphe simple $G = (X, E)$ non pondéré et non orienté. Il ne s'agit pas d'effectuer un partitionnement puisqu'on ne cherche pas forcément à classer tous les sommets et que les classes trouvées peuvent être empiétantes. Le nombre de classes est calculé par l'algorithme et n'est donc pas fixé à l'avance. On n'impose aucune contrainte au départ (partition exacte, nombre de classes, cardinal des classes...) afin d'obtenir des classes « naturelles », intrinsèques au graphe.

Après avoir détaillé les deux étapes de l'algorithme, nous illustrons la méthode par un exemple simple.

Étape 1 : création des noyaux des classes

On considère une fonction de densité locale De évaluant la densité en arêtes au voisinage d'un sommet quelconque (voir plus bas pour l'expression de De). On cherche tous les sommets s réalisant des maxima locaux de cette fonction De et de densité supérieure à la moyenne, c'est-à-dire vérifiant :

$$\forall s' \in \Gamma(s), De(s) \geq De(s') \text{ et } De(s) \geq \overline{De}$$

où $\Gamma(s)$ désigne l'ensemble des sommets adjacents à s et $\overline{De}$ la moyenne des densités sur tous les sommets du graphe : $\overline{De} = \frac{1}{|X|} \sum_{s \in X} De(s)$. Ces maxima

locaux formeront les noyaux initiaux des classes ; si plusieurs de ces sommets sont adjacents, on les place dans le même noyau.

Ensuite on ajoute à chaque noyau tout sommet s qui lui est adjacent et dont la densité $De(s)$ est supérieure ou égale à $\overline{De}$. Si un tel sommet est adjacent à plusieurs noyaux, on ne le classe pas lors de cette étape : les noyaux créés sont disjoints.

pour tout sommet s :
- calculer sa densité locale $De(s)$

calculer la densité moyenne $\overline{De}$

pour tout sommet s :
- déterminer si s est un maximum local, c'est-à-dire s'il vérifie : $\forall s' \in \Gamma(s)$, $De(s) \geq De(s')$

considérer le sous-graphe G_{opt} réduit aux sommets maxima locaux
initialiser les noyaux par les composantes connexes de G_{opt}, lesquelles sont déterminées par un parcours de graphe en profondeur (voir [9])

pour tout sommet s de G non classé et tel que $De(s) \geq \overline{De}$:
- si s possède un ou plusieurs voisins classés, et que ceux-ci sont dans un même noyau, on classe s dans ce noyau
- sinon on ne classe pas s.

Étape 2 : Extension des noyaux

C'est cette étape que nous avons modifiée afin d'engendrer des classes empiétantes. On considère une fonction évaluant la « qualité » (densité en arêtes, cardinal) d'un sous-ensemble de X (voir plus bas pour l'expression de cette fonction). Le principe de l'extension est d'ajouter itérativement à chaque classe les sommets qui y sont connectés s'ils permettent d'augmenter la qualité de la classe (qu'ils soient déjà classés ou non) :

pour chaque classe :
- calculer la valeur de la fonction de qualité
- répéter :
- *pour chaque sommet adjacent à la classe courante, calculer le nombre d'arêtes le reliant à cette classe*
- *sélectionner tous les sommets réalisant le maximum de ce nombre ; ils forment l'ensemble des candidats*
- *calculer la qualité de la classe formée par la classe courante et l'ensemble des sommets candidats*
- *si la valeur calculée est supérieure ou égale à celle de la classe courante :*
  - ◇ *placer les candidats dans la classe*
  - ◇ *mettre à jour la qualité de la classe*

- tant que des sommets ont été ajoutés à cette itération
s'il existe des classes identiques, n'en garder qu'un exemplaire.

Les ensembles de candidats traités ne sont pas nécessairement disjoints puisqu'un sommet peut être adjacent à plusieurs classes, les classes obtenues peuvent donc être empiétantes.

2.1 Fonctions de densité locale

Pour la construction des noyaux des classes, on utilise une fonction de densité locale De donnant la densité en arêtes au voisinage des sommets d'un graphe. T. Colombo *et al* ([7]) proposent et comparent les quatre fonctions de densité suivantes :

- Le degré $d(s)$ de s divisé par le plus grand degré Δ dans le graphe :
$$De_1(s) = \frac{d(s)}{\Delta}.$$
- Le degré moyen dans le voisinage de s : il s'agit de la somme des degrés du sommet s et de ses sommets adjacents divisée par le nombre de sommets considérés :
$$De_2(s) = \frac{d(s) + \sum_{s' \in \Gamma(s)} d(s')}{1 + d(s)}.$$
- Le taux de triangles : on considère le nombre $N_t(s)$ de triangles ayant s pour sommet (c'est-à-dire le nombre d'arêtes reliant entre eux les voisins de s) et on le divise par le nombre maximum de triangles réalisables autour d'un sommet du degré de s :
$$De_3(s) = \frac{N_t(s)}{\frac{1}{2}d(s)(d(s)-1)}.$$
- Le pourcentage d'arêtes dans le voisinage de s : on divise le nombre d'arêtes autour de s (d'extrémité s ou reliant deux sommets adjacents à s) et on le divise par le nombre maximum d'arêtes sur l'ensemble formé par s et ses voisins :
$$De_4(s) = \frac{d(s) + N_t(s)}{\frac{1}{2}d(s)(d(s)+1)}.$$

Ces auteurs montrent par simulation sur des graphes aléatoires dans lesquels figurent des zones de densité plus forte que la moyenne que les fonctions De_3 et De_4 sont plus satisfaisantes que De_1 et De_2. Nous allons utiliser ici la fonction De_4, à valeurs dans $[0\,;1]$, qui permet le mieux de retrouver ces zones. Nous imposons qu'un sommet de degré 1 soit de densité nulle afin de ne pas former de noyaux à partir de tels sommets.

2.2 Critères d'extension

Il s'agit de définir une notion de qualité d'un sous-ensemble de sommets. On autorisera alors l'ajout d'un sommet dans une classe si cela augmente cette qualité. La fonction de qualité doit à la fois dépendre :

- du pourcentage d'arêtes du sous-graphe, le but étant de mettre en évidence des zones denses en arêtes ;
- de l'ordre du sous-ensemble considéré. La fonction ne doit pas être nécessairement maximale pour des cliques ; à densité égale, une classe d'ordre élevé doit avoir une qualité plus grande qu'une classe d'ordre inférieur.

Nous proposons dans cette section deux critères vérifiant ces propriétés.

On notera H un sous-graphe de G, p son ordre (nombre de sommets) et q sa taille (nombre d'arêtes). On notera S l'ensemble des sommets candidats à entrer dans la classe H ; ce sont les sommets les plus connectés à la classe, c'est-à-dire reliés chacun par c arêtes aux sommets de la classe, c étant le nombre d'arêtes maximum entre un sommet extérieur à la classe et les sommets de la classe.

Critère du degré moyen

Étant données les remarques précédentes, il semble immédiat de considérer le degré moyen comme fonction de qualité :

$$d_m(H) = \frac{\sum_{s \in H} d(s)}{p} = \frac{2.q}{p}.$$

En effet plus un sous-graphe est dense en arêtes, plus son degré moyen est élevé et un petit sous-graphe aura en moyenne un degré moyen moins élevé qu'un gros sous-graphe, même si c'est une clique ($d_m(H) \leq p - 1$).

Si on considère une classe H qu'on cherche à étendre et S l'ensemble des candidats, la règle d'extension est alors :

$$H \longleftarrow H \cup S \text{ ssi } d_m(H \cup S) \geq d_m(H).$$

Lorsque l'ensemble S est réduit à un singleton, cette extension revient à ajouter à la classe le candidat si $c \geq \frac{q}{p}$ (on rappelle que c est le nombre d'arêtes entre chaque sommet de S et les sommets de la classe), ce qui s'applique en particulier aux cliques qu'on pourra donc étendre suivant ce critère.

On peut, plus généralement, s'intéresser aux règles d'extension suivantes :

$$H \longleftarrow H \cup S \text{ ssi } c \geq \frac{\alpha.q}{p}$$

où α est un nombre réel strictement positif. Plus le coefficient α est élevé et plus le critère d'extension est strict : si α est petit (tel que $\frac{\alpha.q}{p} < 1$), la classe sera étendue par tous ses sommets adjacents ; s'il est grand ($\frac{\alpha.q}{p} > \Delta$), aucune extension n'aura lieu.

On obtient alors une famille de règles d'extension (critère du degré moyen généralisé) plus ou moins strictes suivant la valeur de α. On pourra déterminer cette dernière en fonction du graphe traité et des exigences sur les classes (densité, cardinal).

Critère probabiliste

On constate que les zones denses en arêtes sont rares (surtout dans un graphe ayant une densité en arêtes relativement faible, ce qui est le cas dans l'application

envisagée plus loin). On propose alors comme critère d'extension d'étendre une classe si son extension est moins probable que la classe de départ.

Pour chaque classe H (sous-graphe de G) ayant p sommets, considérons la probabilité $P(H)$ que H ait q arêtes sachant que le graphe initial G possède m arêtes pour n sommets. On étendra H si les sommets ajoutés diminuent la valeur de cette probabilité.

Soit $M = \frac{n(n-1)}{2}$ le nombre maximum de paires qu'on peut former avec les sommets de G et $Q = \frac{p(p-1)}{2}$ le nombre maximum de paires qu'on peut former avec ceux de H.

Proposition 1. *La probabilité $P(H)$ vaut (loi hypergéométrique) :*

$$P(H) = \frac{C_m^q C_{M-m}^{Q-q}}{C_M^Q}.$$

Démonstration. Sur ces M paires, il y en a m qui représentent une arête du graphe (sommets adjacents). On cherche la probabilité que H contienne exactement q paires de sommets adjacents de G. Cette probabilité est égale à celle de tirer exactement q éléments possédant une certaine propriété (ici, de définir une arête) lors de Q tirages sans remise dans un ensemble de M éléments dont m possèdent cette propriété. D'où le résultat.

Si on considère une classe H qu'on cherche à étendre et l'ensemble S des candidats, la règle d'extension est :

$$H \longleftarrow H \cup S \text{ ssi } P(H \cup S) \leq P(H).$$

2.3 Exemple

Nous traitons dans cette partie un exemple en appliquant pas à pas notre algorithme au graphe G (8 sommets, 12 arêtes) représenté figure 1.

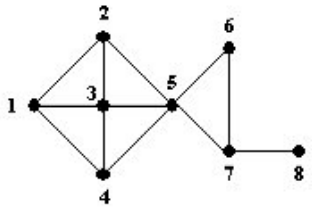

Fig. 1. Graphe G

Étape 1 : création des noyaux des classes

sommet : s	1	2	3	4	5	6	7	8	moyenne
degré : d(s)	3	3	4	3	5	2	3	1	3
nb de triangles : $N_t(s)$	2	2	4	2	3	1	1	0	
densité : $De_4(s)$	0,833	0,833	0,8	0,833	0,533	1	0,666	0	0,687

Tab. 1.

On commence par calculer le pourcentage d'arêtes au voisinage de chaque sommet : $De_4(s) = \frac{d(s)+N_t(s)}{\frac{1}{2}d(x)(d(x)+1)}$ (voir tableau 1).

Les optima locaux pour cette fonction sont les sommets 1, 2, 4 et 6. Les sommets 1, 2 et 4 forment un sous graphe connexe, ils sont donc placés dans le même noyau ; 6 forme le deuxième noyau. On examine ensuite les voisins de ces noyaux. Le sommet 3 est de densité supérieure à la moyenne (0,687) et adjacent au noyau 1 : on le place dans ce noyau. Le deuxième noyau n'a pas de voisin de densité supérieure à la moyenne ; il n'est donc pas étendu lors de cette étape.

- Noyau 1 : $\{1, 2, 3, 4\}$
- Noyau 2 : $\{6\}$

Étape 2 : Extension des classes

Nous allons utiliser le critère du degré moyen généralisé avec $\alpha = 2$. Commençons par étendre le noyau 1 :

- classe 1 : $\{1, 2, 3, 4\}$, $\frac{2q}{p} = \frac{10}{4} = 2,5$.

On détermine l'ensemble S des sommets candidats à entrer dans la classe ; ici le sommet 5 est le seul adjacent au noyau : $S = \{5\}$. Il est relié par 3 arêtes au noyau : $c = 3$. On vérifie si $c \geq \frac{2q}{p}$. Ici c'est le cas, on ajoute donc le sommet 5 à la classe.

- classe 1 : $\{1, 2, 3, 4, 5\}$, $\frac{2q}{p} = \frac{16}{5} = 3,2$.

On détermine à nouveau l'ensemble des candidats : $S = \{6, 7\}$, $c = 1$. L'inégalité $c \geq \frac{2q}{p}$ n'est pas vérifiée, la classe n'est donc pas étendue et on s'arrête. Étendons maintenant le noyau 2 :

- classe 2 : $\{6\}$, $\frac{2q}{p} = \frac{0}{1} = 0$.

Les candidats sont $S = \{5, 7\}$, avec $c = 1$. Puisque $c \geq \frac{2q}{p}$, on étend la classe en y ajoutant ces sommets.

- classe 2 : $\{5, 6, 7\}$, $\frac{2q}{p} = \frac{6}{3} = 2$.

Les sommets adjacents à la classe sont 2, 3, 4 et 8. Ils sont tous candidats car tous reliés par une arête à la classe : $S = \{2, 3, 4, 8\}$, $c = 1$. Le critère d'extension n'est pas vérifié donc on n'étend pas la classe et on s'arrête.

L'algorithme a construit deux classes : $\{1, 2, 3, 4, 5\}$ et $\{5, 6, 7\}$, résultat qui semble intuitivement satisfaisant. Il y a un sommet non classé (8) et un sommet classé dans deux classes (5). L'utilisation du critère d'extension probabiliste donne pour ce graphe les mêmes classes.

3 Application à un graphe d'interactions entre protéines

Notre travail est réalisé dans le cadre d'une étude interdisciplinaire avec une équipe de biologie moléculaire (B. Jacq *et al*, LGPD, Marseille) qui nous a fourni un graphe d'interactions protéine-protéine pour la levure de 1777 sommets et 2630 arêtes qui correspondent à des contacts. On cherche à déterminer les zones denses en arêtes de ce graphe car, selon une hypothèse biologiste fortement défendable ([4]), il est vraisemblable que deux protéines ayant beaucoup d'interacteurs communs contribuent à une même fonction cellulaire. On cherche de plus à obtenir des classes interprétables par les biologistes (de cardinal compris environ entre 5 et 25). Nous allons étudier et comparer expérimentalement les différents critères d'extension.

Dans ce graphe, très peu dense (densité de 0,0017) et peu homogène, notre algorithme construit 467 noyaux initiaux et 706 sommets sont classés initialement dans les noyaux.

3.1 Critère du degré moyen généralisé

Tout d'abord, comparons le critère du degré moyen généralisé pour différentes valeurs de α. On rappelle que suivant ce critère, on ajoute à la classe H (possédant p sommets et q arêtes) l'ensemble S des sommets qui lui sont le plus connectés (par c arêtes pour chaque sommet de S) de la façon suivante :

$$H \longleftarrow H \cup S \text{ ssi } c \geq \frac{\alpha.q}{p}.$$

Nous avons traité le graphe d'interactions pour α variant dans [0,5 ; 5]. La figure 2 présente les résultats obtenus : nous avons calculé le pourcentage d'éléments classés (graphique 1), le nombre de classes (graphique 2), la moyenne des degrés moyens des classes (graphique 3), la moyenne des densités internes des classes (graphique 4), la moyenne des cardinaux des classes (graphique 5) et le nombre moyen de classes par sommet (graphique 6). Les pointillés des graphiques 2 et 3 représentent respectivement le nombre de classes à l'issue de l'étape 1 et le degré moyen du graphe.

Nous remarquons tout d'abord que si α augmente (c'est-à-dire si le critère d'extension devient plus strict), il y a de moins en moins d'éléments classés, les classes deviennent de plus en plus petites mais de plus en plus denses. Il est en effet intuitif que plus une classe est petite, plus il est « facile » qu'elle soit dense (une classe à deux sommets sera automatiquement de densité 1). La courbe du degré moyen est la seule qui ne soit pas monotone, elle croît, atteint son maximum en $\alpha = 0,8$ puis décroît. En effet le degré moyen des classes dépend en partie de leur cardinal puisqu'il est majoré par $p - 1$ (une classe à deux sommets ne peut avoir un degré moyen supérieur à 1), il décroît donc quand α augmente. Quand α est suffisamment petit, le cardinal des classes est grand et cette majoration n'a plus d'effet : on se rapproche du degré moyen du graphe

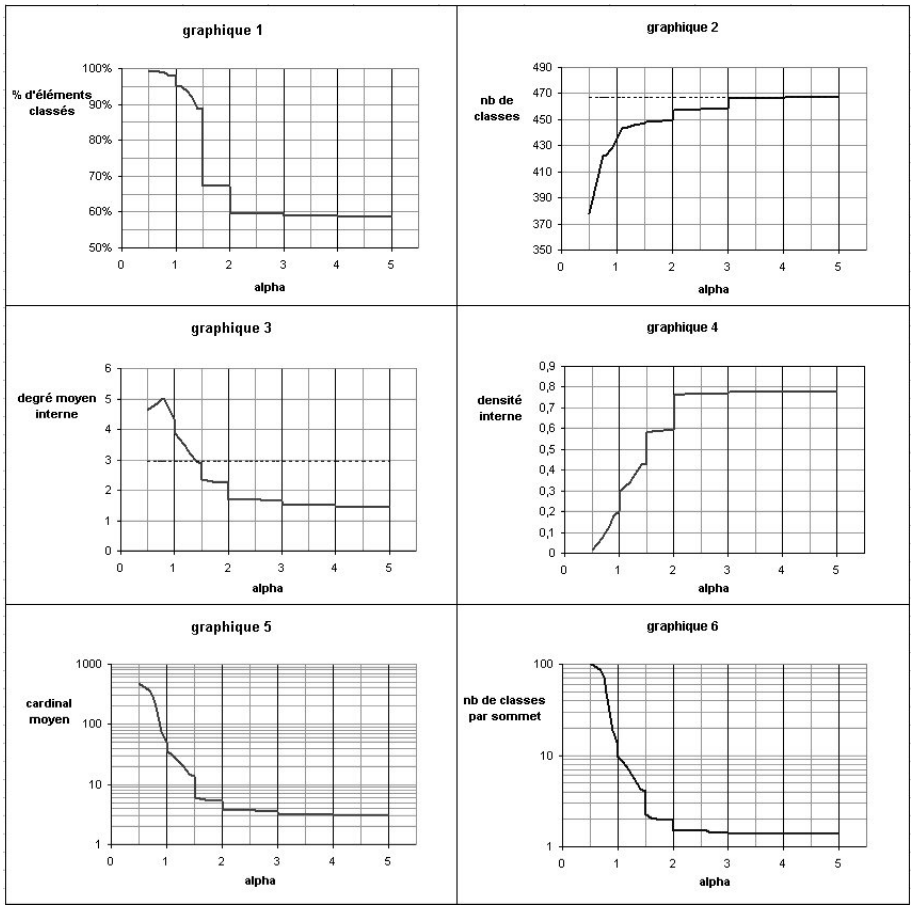

Fig. 2. Résultats obtenus par la méthode du degré moyen suivant les valeurs de α

entier. On peut choisir la valeur de α réalisant le maximum du degré moyen des classes obtenues mais elle correspond à des classes de cardinal moyen 194 et qui ne sont donc pas facilement interprétables biologiquement. On observe que les valeurs de α pertinentes pour tous les critères simultanément sont dans $[1; 2]$, on choisit ensuite α suivant qu'on veut plutôt privilégier la densité des classes, leur degré moyen ou leur taille.

La seconde remarque sur ces graphes concerne la discontinuité des courbes en $\alpha = 1\,;\, 1,5\,;\, 2\,;\, 3$ ... et les paliers relativement constants qui les séparent. Ceci s'explique lorsque l'on observe ce qui se passe lors de l'extension des noyaux.

Considérons les configurations initiales suivantes :

p=2
q=1

$$H \longleftarrow H \cup S \text{ ssi } c \geq \frac{\alpha}{2}.$$

- si $\alpha \leq 2$, on peut ajouter au noyau tout sommet qui lui est adjacent.
- si $2 < \alpha \leq 4$, on lui ajoute les sommets vérifiant $c \geq 2$.
- si $\alpha > 4$ on ne peut étendre la classe puisque $c \leq p = 2$.

p=3
q=2

$$H \longleftarrow H \cup S \text{ ssi } c \geq \frac{2\alpha}{3}.$$

- si $\alpha \leq \frac{3}{2}$, on peut ajouter au noyau tout sommet qui lui est adjacent.
- si $\frac{3}{2} < \alpha \leq 3$, on lui ajoute les sommets vérifiant $c \geq 2$.
- si $3 < \alpha \leq \frac{9}{2}$, on lui ajoute les sommets vérifiant $c \geq 3$.
- si $\alpha > \frac{9}{2}$ on ne peut étendre la classe puisque $c \leq p = 3$.

p=3
q=3

$$H \longleftarrow H \cup S \text{ ssi } c \geq \alpha.$$

- si $\alpha \leq 1$, on peut ajouter au noyau tout sommet qui lui est adjacent.
- si $1 < \alpha \leq 2$, on lui ajoute les sommets vérifiant $c \geq 2$.
- si $2 < \alpha \leq 3$, on lui ajoute les sommets vérifiant $c \geq 3$.
- si $\alpha > \frac{9}{2}$ on ne peut étendre la classe puisque $c \leq p = 3$.

En superposant les comportements de l'extension de ces trois configurations initiales on retrouve l'allure des courbes de résultats (discontinuité puis palier). De nombreuses autres valeurs de α jouent un rôle, les courbes ne sont donc pas tout à fait constantes entre ces points de discontinuités.

On observe que pour ce graphe les résultats sont très satisfaisants puisqu'une grande partie des sommets sont classés (de 59 à 99 % suivant les valeurs de α considérées), que la densité des classes obtenues est très largement supérieure à la densité moyenne du graphe (pour $\alpha = 2$, la densité interne est de 0,59, ce qui correspond à peu près à 350 fois la moyenne du graphe), et que le degré moyen des classes reste relativement élevé (toujours pour $\alpha = 2$, il est de 2,23 alors que le degré moyen du graphe est de 2,96).

Néanmoins, si on s'intéresse aux cardinaux des classes obtenues, dont la moyenne est relativement faible (du moins pour α assez grand), on constate qu'il existe beaucoup de classes de cardinal très petit (pour $\alpha = 2$, plus de la moitié des classes sont formées de 2 ou 3 éléments) qui ne sont pas tellement intéressantes mais qui contribuent à augmenter la densité interne moyenne. La figure 3 représente la répartition des classes en fonction de leurs cardinaux (pourcentages cumulés) pour $\alpha = 1$, $\alpha = 1,5$ et $\alpha = 2$.

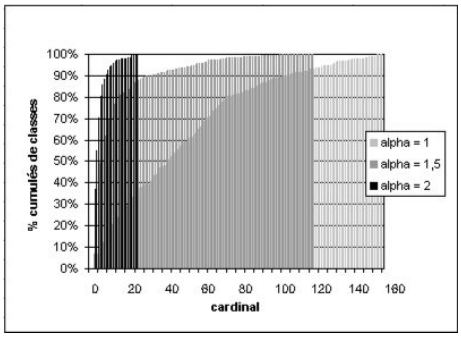

Fig. 3. Répartition des classes par cardinaux

Il n'est donc pas forcément intéressant de choisir α trop grand car alors peu de classes ont un cardinal élevé.

En ce qui concerne l'empiètement des classes, on remarque (figure 2, graphique 6) que de nombreux sommets sont classés dans plusieurs classes (le nombre de classes par sommets est très grand pour α inférieur à 1, compris entre 2 et 10 pour $\alpha \in [1;2]$ et même avec α élevé il reste strictement supérieur à 1).

3.2 Critère d'extension probabiliste

Nous allons maintenant comparer cette méthode avec la méthode d'extension probabiliste. Suivant ce critère, on étend la classe H par l'ensemble des sommets candidats S si $P(H \cup S) \leq P(H)$. Les résultats obtenus sont présentés dans la première ligne du tableau 2. La seconde ligne du tableau contient les résultats obtenus avec la méthode du degré moyen pour $\alpha = 1,22$ (valeur choisie de manière à avoir le même cardinal moyen).

méthode	nb de classes	% d'éléments classés	nb de classes par sommet	card moy	dens moy	deg moy
proba	430	93,25%	6,9	23,9	0,22	3,19
$\alpha = 1,22$	444	93,19%	6,7	23,9	0,35	3,4

Tab. 2.

D'après ces valeurs, la méthode du degré moyen semble significativement plus efficace puisqu'il y a à peu près autant d'éléments classés, que les classes formées sont plus denses et ont un degré moyen plus élevé. On va maintenant étudier la répartition des classes par cardinaux dans chacun des cas (figure 4).

On remarque que les cardinaux des classes sont répartis de façon beaucoup plus équilibrée autour de la moyenne par la méthode d'extension probabiliste.

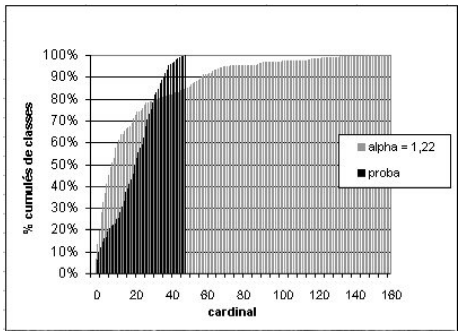

Fig. 4. Répartition des classes par cardinaux

La médiane pour celle-ci est entre 24 et 25 alors que pour la méthode du degré moyen généralisé ($\alpha = 1,22$) elle se situe entre 11 et 12. La méthode d'extension probabiliste est donc à considérer si on souhaite obtenir un plus grand nombre de classes de taille intermédiaire (ce qui peut être souhaitable pour faciliter l'interprétation biologique des classes). On peut aussi souligner que les résultats pour la densité interne (130 fois la densité moyenne du graphe) et le degré moyen interne sont toujours très satisfaisants.

3.3 Validation de la méthode

Nous disposons des résultats obtenus par une autre méthode de partitionnement appliquée à d'autres données d'interactions toujours chez la levure ([5]). Cette méthode n'opère pas directement sur le graphe; elle produit des classes non empiétantes qui ont été validées d'un point de vue biologique. Nous appellerons donc ces classes *classes de référence*. Les données sont ici un graphe de 877 protéines d'une densité de 0,005. Nous avons traité ces données par notre programme et comparé nos classes avec les classes de référence afin de valider notre méthode.

méthode	nb de classes	% d'élts classés	card moy	dens moy	deg moy
référence	126	100 %	6,9	0,45	1,95
proba	250	96,9%	25,7	0,2	3,87
$\alpha = 1$	186	97,1%	54.4	0,19	4,62
$\alpha = 1,5$	238	94,5%	11,6	0,45	3,03
$\alpha = 2$	269	89,2%	5,6	0,6	2,45

Tab. 3.

Le tableau 3 donne les caractéristiques des classes de référence et de celles trouvées par notre programme pour les différents critères d'extension. Tout d'abord ces résultats confirment les conclusions tirées du premier graphe sur

notre méthode : les valeurs de α intéressantes sont ici aussi comprises dans $[1;2]$, au-delà les classes trouvées ne sont pas intéressantes car de cardinal trop élévé ou trop faible. Nous avons aussi obtenu les mêmes allures des courbes de répartition des classes par cardinal, la méthode d'extension probabiliste donnant toujours des classes plus uniformes en terme de cardinaux.

Pour chaque classe de référence et pour chaque méthode d'extension, nous avons déterminé la classe lui correspondant parmi les classes données par la méthode, c'est-à-dire la classe ayant le plus d'éléments communs avec elle. Nous calculons ensuite le pourcentage d'éléments de la classe de référence appartenant à sa classe correspondante. Nous avons aussi calculé le pourcentage de paires d'éléments dans une méme classe de référence qui sont aussi classées ensemble par notre méthode. Les résultats trouvés sont donnés dans le tableau 4.

methode	% d'éléments bien classés	% de paires conservées
proba	89 %	77%
$\alpha = 1$	87%	81%
$\alpha = 1,5$	75%	57%
$\alpha = 2$	65%	21%

Tab. 4.

La méthode d'extension probabiliste et celle du degré moyen pour $\alpha = 1$ conservent une très grande proportion des classes de référence puisque environ 25% des classes de références sont incluses dans leur classe correspondante et 55% ont une intersection avec leur classe correspondante représentant plus de 90%. Pour $\alpha = 1,5$ cette proportion reste élevée. Pour $\alpha = 2$ les résultats ne sont pas si bons car les classes trouvées sont en moyenne de cardinal inférieur aux classes de référence. Si dans ce cas-là on considère le pourcentage par rapport aux classes trouvées, on trouve qu'en moyenne 70% des éléments des classes sont dans une même classe de référence.

Nous avons ainsi par notre programme retrouvé en grande partie les classes de référence, et obtenu des classes de cardinal variable qui permettront peut-être de nouvelles prédictions fonctionnelles.

4 Conclusion

D'un point de vue biologique, les classes disjointes n'étaient pas fondées, il était donc essentiel d'étudier la possibilité de chevauchement des classes. Cette méthode de classification empiétante donne des résultats satisfaisants lorsqu'on l'applique à des graphes d'interactions entre protéines. Les différents critères d'extension proposés permettent à l'utilisateur de privilégier certains caractères des classes obtenues. Le paramètre α permet de construire des classes de cardinalité variable : connaissant la taille moyenne d'un complexe biologique, on pourra ajuster la valeur de α de façon à pouvoir envisager les classes d'un point de vue fonctionnel. On pourra aussi utiliser le critère d'extension probabiliste si

on souhaite obtenir des classes de taille plus homogène. Reste à nos collègues biologistes d'analyser les classes obtenues et de déterminer quel critère est le mieux adapté.

5 Remerciements

Nous remercions l'ACI IMPBIO et l'équipe de bio-informatique du LGPD (CNRS, Marseille) qui nous a fourni les données d'interaction des protéines de la levure.

Références

1. P. Arabie, L. J. Hubert, G. De Soete, Clustering and Classification, World Scientific, Singapore, New Jersey, London, Hong Kong, 1996.
2. J.-P. Barthélemy, G. Cohen, A. Lobstein, Complexité algorithmique et problèmes de communications, Masson, Paris, 1992
3. G. Brossier, Les éléments fondamentaux de la classification, in Analyse des données, G. Govaert (éd.), Hermès Lavoisier, Paris, 2003, 235-262.
4. C. Brun, J.Wojcik, A. Guénoche, B. Jacq, Étude bioinformatique des réseaux d'interactions : PRODISTIN, une nouvelle méthode de classification des protéines, acte de JOBIM 2002, Saint-Malo, 171-182.
5. C. Brun, C. Herrmann, A. Guénoche, Clustering proteins from interaction networks for the prediction of cellular functions, BMC Bioinformatics, 2004, 5 :95.
6. J.L. Chandon, S. Pinson, Analyse topologique - Théories et applications, Masson, Paris, 1981.
7. T. Colombo, Y. Quentin, A. Guénoche, Looking for high density areas in a graph : application to orthologous genes, soumis pour publication.
8. T. Colombo, Y. Quentin, A. Guénoche, Recherche de zones denses dans un graphe : application aux gènes orthologues, Colloque « Knowledge Discovery and Discrete Mathematics», Actes des Journées Informatiques de Metz, INRIA, 2003, 203-212.
9. T. Cormen, C. Leiserson, R. Rivest, Introduction à l'algorithmique, Dunod, Paris, 1994.
10. W.H.E. Day, Complexity theory : An introduction for practitioners of classification, in Clustering and Classification, P. Arabie, L. J. Hubert, G. De Soete (éds.), World Scientific, Singapore, New Jersey, London, Hong Kong, 1996, 199-233.
11. M.R. Garey, D.S. Johnson, Computers and intractability, a guide to the theory of NP-completeness, Freeman, New York, 1979.
12. P. Hansen, B. Jaumard, Cluster analysis and mathematical programming, Mathematical Programming 79, 1997, 191-215.
13. D. Wishart, Mode analysis : generalization of nearest neighbour which reduces chaining effects, *Numerical taxonomy*, academic press, 1976, 282-311.