Décomposition homogène des réseaux macromoléculaires

Del Mondo Géraldine

encadrée par Eveillard Damien et Rusu Irena

Laboratoire d'Informatique de Nantes Atlantique
2, rue de la Houssinière
B.P. 92208
F-44322 NANTES CEDEX 3

Rapport de Stage de Master Recherche

Septembre 2007

LINA, Faculté des Sciences & Techniques de Nantes
2, rue de la Houssinière
B.P. 92208 — F-44322 NANTES CEDEX 3

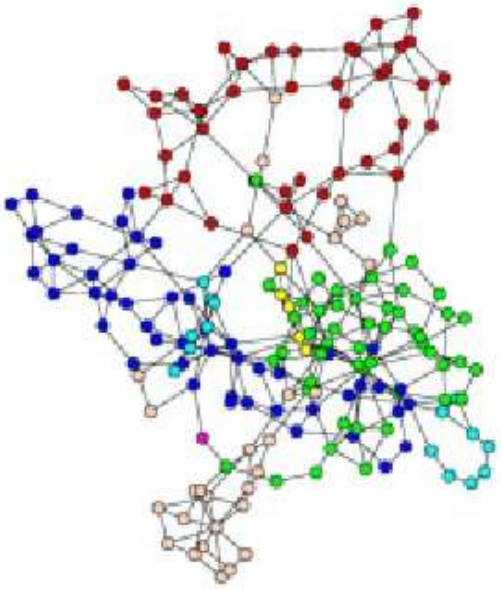
Del Mondo Géraldine (encadrée par Eveillard Damien et Rusu Irena)

Décomposition homogène des réseaux macromoléculaires

Décomposition homogène des réseaux macromoléculaires

Del Mondo Géraldine (encadrée par Eveillard Damien et Rusu Irena)

geraldine.del-mondo@etu.univ-nantes.fr



Remerciements

Merci, merci à vous qui avez dû me supporter, m'entourer, m'aider.

Vous, mes parents.

Vous, mes professeurs, et en particulier Damien Eveillard et Irena Rusu mes deux encadrants qui m'ont fait confiance en me prenant en stage.

Vous, mes amis, en particulier Jimmy pour ses remarques scientifiques avisées.

Vous, chercheurs qui pour certains m'ont fourni des données et/ou programme (Julien Gagneur, Roger Guimera, Hongwu Ma, Tatsuo Shibata...). Je remercie en particulier Julien Gagneur pour m'avoir fourni son implémentation de la décomposition modulaire et pour les discussions que nous avons pu avoir par mail.

J'espère que tous, vous serez satisfaits du travail accompli ;-)

Motivations

La modularité dans les systèmes vivants La biologie est l'étude des systèmes vivants. Les systèmes vivants sont par essence des systèmes complexes et leur compréhension passe par la décomplexification de la machinerie biologique [Barabási and Oltvai, 2004]. L'un des moyens d'y parvenir est de décomposer le système en petits morceaux interprétables en terme de fonctionnalités. Cet aspect modulaire [Szallasi et al., 2001] [Rives and Galitski, 2003] est globalement aujourd'hui accepté par la communauté scientifique.

L'objectif étant de déterminer les différents modules d'une partie d'un système biologique, les réseaux d'interactions de protéines sont modélisés sous forme de graphe. Les protéines sont modélisées par les sommets, et une arête entre deux protéines implique l'existence d'une interaction entre ces deux protéines. Ces interactions forment des réseaux macromoléculaires complexes qu'il faut analyser de façon efficace, notamment à l'aide d'une décomposition. La mise en oeuvre de "diviser pour régner" est un concept efficace en théorie des graphes ce qui explique l'intérêt de la décomposition de graphes.

Néanmoins, la description d'un module biologique n'est pas précisément définie, compliquant sa recherche automatique. De récentes études ([Holme et al., 2003], [Ishihara et al., 2005], [Zotenko et al., 2006], [Adamcsek et al., 2006], [Guimera and Amaral, 2005]...), considèrent un module comme une unité fonctionnelle semi-autonome composée d'éléments qui sont connectés de façon plus forte entre eux qu'avec le reste de l'environnement. [Gagneur et al., 2004] s'appuie par ailleurs sur une autre définition usuelle et plus contrainte du module issue de la théorie des graphes (voir section 1.1).

Méthodes de décomposition des réseaux biologiques Il existe différentes méthodes pour décomposer les graphes biologiques.

- [Ravasz and al, 2002] utilise une matrice topologique de chevauchement[1], la direction des réactions est ignorée et seule la connectivité locale (degré de connexion des sommets du réseau) est considérée. Ils classent les sommets dans différents groupes, mais d'un point de vue biologique, un sous ensemble de sommets n'est pas suffisant pour définir un unique *chemin fonctionnel*[2] ou module fonctionnel. Il doit aussi y avoir les réactions entre ces métabolites dans ce sous ensemble.
- [Schuster et al., 2002] enlève les métabolites les plus fortement connectées. L'idée est d'étiqueter les métabolites avec un degré k dans le graphe, supérieur à un seuil k_{max} dit "ex-

[1] Topological overlap matrix : matrice d'adjacence dans laquelle on place un 1 en cas de noeuds connectés et 0 sinon. Une mesure (topological overlap) est calculée par paire de noeuds afin de savoir dans quelle mesure ils partagent les mêmes voisins.

[2] Séquence de réactions enzymatiques ou autres, par lesquelles un matériel biologique est transformé en un autre.

térieur" et considérer les composantes connexes composées des métabolites "intérieures" comme des sous réseaux. La méthode détecte des sous réseaux, mais certains réseaux peuvent ne pas avoir de propriétés locales détectables par des quantités locales comme les degrés. Cette méthode est néanmoins capable de décrire l'organisation hiérarchique complète d'un réseau cellulaire, i.e un réseau macromoléculaire divisé en sous réseaux et ces sous réseaux en sous réseaux etc...

- [Gagneur et al., 2004] étend la méthode de [Schuster et al., 2002] en groupant les réactions selon le degré de connexion des métabolites connectés afin d'organiser la structure hiérarchique. Son implémentation issue d'une méthode appelée décomposition modulaire[Gallaï, 1967] (voir section 1) mène à un arbre $T(G)$ où les feuilles sont les sommets du graphe G de départ, et où les noeuds internes correspondent à certaines opérations sur les graphes. Cependant, la caractérisation choisie (degré de connection) par définition locale, ne suffit pas toujours pour décomposer le réseau. Il faudrait une caractérisation prenant en compte l'organisation globale du réseau.
- Pour résoudre ce même problème ([Schuster et al., 2002], [Gagneur et al., 2004]), [Holme et al., 2003] développent une méthode qui révèle la hiérarchie des sous-réseaux du réseau global. Elle utilise une combinaison des propriétés locales et globales du réseau métabolique[3]. Le degré de connection des métabolites est une propriété structurelle locale du réseau. Leur méthode de décomposition est basée sur la structure *bow-tie* (noeud papillon) et la longueur du plus court chemin entre réactions. Ces deux notions reflètent la connectivité globale de la structure du réseau. Les sous-réseaux ont une taille adéquate par rapport à des unités fonctionnelles et correspondent à une fonction biologique. Ils peuvent servir de base à une analyse fonctionnelle du réseau métabolique.

Un exemple de méthode ad-hoc de décomposition [Guimera and Amaral, 2005] proposent une méthode pour extraire l'information contenue dans des réseaux complexes basée sur la connectivité des noeuds. L'idée est que les noeuds ayant un même rôle doivent avoir les mêmes propriétés topologiques. Il est alors possible d'identifier des modules fonctionnels dans ces réseaux et de classer les noeuds selon des rôles "universels" selon leurs connections inter et intra modules. Cette méthode sera testée en section 5.3.

Une nouvelle voie : la décomposition homogène Après l'étude des résultats obtenus par les différentes décompositions, nous nous proposons d'améliorer l'approche de [Gagneur et al., 2004]. Notre finalité est de décomposer au maximum un réseau biologique. Nous utiliserons la décomposition homogène [Jamison and Olariu, 1995] qui étend la décomposition modulaire (voir section 2).

A partir du graphe G modélisant le réseau, nous allons calculer l'arbre de décomposition modulaire $T_M(G)$ (abrégé T_M si pas d'ambiguïté). De cet arbre que nous obtiendrons l'arbre de décomposition homogène de G, $T_H(G)$ (abrégé T_H si pas d'ambiguïté). (voir la section 3)

Cette démarche nous permettra de compléter les informations que peut fournir la décomposition modulaire (section 5) mais aussi de comparer nos résultats in silico avec les modules fonctionnels biologiques in vivo.

[3]Suite de réactions chimiques contrôlées éventuellement par des enzymes, leur but est d'exploiter et de transformer les ressources disponibles en énergie (définition disponible : http ://www.irisa.fr/videos/EcoleChercheurBioInfo/siegel/siegel.pdf)

Chapitre 1

La décomposition modulaire

Dans la suite de ce document nous considérons des graphes simples non orientés et sans boucle $G = (V, E)$ où V est l'ensemble des sommets et E l'ensemble des arêtes.

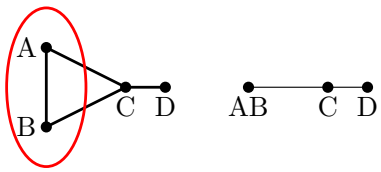
Nous introduisons la décomposition modulaire, définie par [Gallaï, 1967].

1.1 Définitions

Définition 1 *Module*

*$M \subset V$ est un **module** si tous les sommets de M ont les mêmes voisins dans $V - M$.*

Ainsi on peut contracter tous les sommets d'un module car ils ont le même comportement par rapport à l'extérieur[1].(voir figure 1.1)



a) *b)*

FIG. 1.1: Contraction des sommets d'un module en un point. a) A et B ont le même voisin C à l'extérieur du module composé de A et de B. b) On peut donc remplacer A et B par un seul sommet AB.

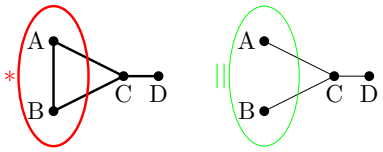
.

On peut définir trois types de modules distincts dont les deux suivants :

- Les modules **séries** figure 1.2a) :
  Les sommets d'un module série forment une **clique**, c'est-à-dire qu'ils sont tous voisins les uns des autres.
- Les modules **parallèles** figure 1.2b) :

[1] L'extérieur d'un module correspond à l'ensemble des sommets qui n'appartiennent pas au module, un même comportement par rapport à l'extérieur signifie que tous les sommets du module ont les mêmes voisins pris dans cet ensemble.

Les sommets d'un module parallèle forment un **stable**, c'est-à-dire qu'ils ne sont pas voisins les uns des autres.



a) *b)*

FIG. 1.2: Deux types de modules distincts, A et B forment un module car ils ont le même voisin C mais : a) A et B sont adjacents donc il s'agit d'un module série, b) A et B sont non adjacents donc il s'agit d'un module parallèle

Tous les modules non triviaux, c'est-à-dire ceux qui ne contiennent ni un seul sommet ni l'ensemble des sommets du graphe, sont appelés ensembles homogènes comme le décrit formellement la définition 2.

Définition 2 *Ensemble homogène*
*$H \subset V$ est un **ensemble homogène** si c'est un module avec $1 < |H| < |V|$.*
Un ensemble homogène est maximal si aucun ensemble homogène ne le contient.

Remarque 1 *Autrement dit, H est un ensemble homogène si tout sommet à l'extérieur de H est adjacent soit à tous sommets de H soit à aucun et si H possède au moins deux sommets.*
Tout ensemble homogène est un module, mais tout module n'est pas un ensemble homogène.
On peut dire qu'un ensemble homogène est un module non trivial.

La définition 2 permet de définir un type de graphe particulier et prépondérant en biologie :

Définition 3 *Graphe premier*
Un graphe sans ensemble homogène est appelé graphe premier ou "prime".

Le **module de type premier** correspondant à ce type de graphe est introduit dans le troisième point du théorème de la décomposition modulaire [Gallaï, 1967].

Théorème 1 *Soit $G = (V, E)$ un graphe quelconque où $|V| \geq 2$, exactement une seule de ces assertions est satisfaite :*

1. *G est non connexe. Dans ce cas, V est un module parallèle et l'ensemble des sommets de chaque composante connexe de G définit un sous-module.*
2. *$\overline{G}$ est non connexe. Dans ce cas, V est un module série et l'ensemble des sommets de chaque composante connexe de $\overline{G}$ définit un sous-module.*
3. *Il existe $Y \subseteq V$, $|Y| \geq 4$ et une unique partition P de V telle que le graphe induit par Y soit un sous-graphe maximal premier de G et pour tout $S \in P$, $|S \cap Y| = 1$.*

De ce théorème on déduit l'arbre de décomposition modulaire associé au graphe G avec les propriétés suivantes :

- Les modules triviaux composés d'un sommet et composés de l'ensemble des sommets V correspondent respectivement aux feuilles et à la racine de l'arbre.

- [2]Un noeud correspondant au module M' est fils de celui correspondant au module M si et seulement si M ne chevauche aucun autre module et $M' \subset M$.
- Il existe trois type de noeud, respectivement selon les cas 1, 2 et 3 du théorème :
  - Un noeud interne étiqueté parallèle (souvent noté || ou 0) signifie que les descendants directs de ce noeud ne sont pas voisins les uns des autres dans le graphe G.
  - Un noeud interne étiqueté série (souvent noté * ou 1) signifie que les descendants directs de ce noeud sont tous voisins les uns des autres dans le graphe G.
  - Pour un noeud étiqueté premier (souvent noté P ou N), les liens qui unissent ses descendants dans le graphe G ne sont pas stockés dans l'arbre.
    Le plus petit graphe premier est le chemin P_4 (on appelle P_k le graphe représentant un chemin sans corde à k sommets). Ci-dessous la figure 1.3 montre un P_4 et son arbre de décomposition modulaire associé dans lequel il y a bien un noeud premier P.

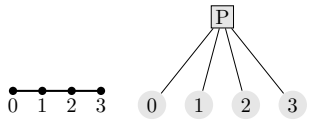


FIG. 1.3: Un P_4 et son arbre de décomposition modulaire

1.2 Exemple de mise en oeuvre de la décomposition modulaire

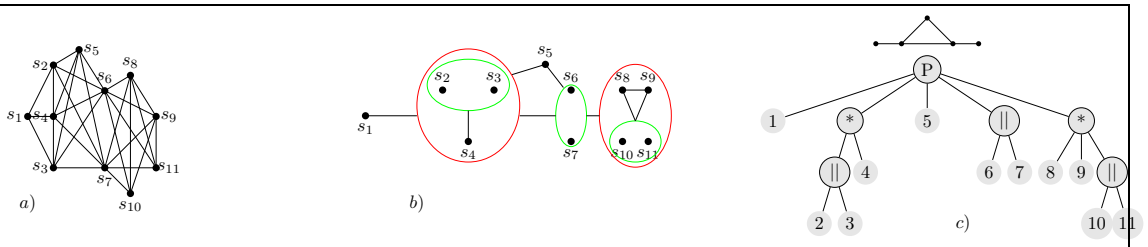


FIG. 1.4: Décomposition modulaire d'un graphe G a) Graphe G. b) Graphe caractéristique détaillé de G : on représente les modules de G et les liens qui existent entre eux. c) Arbre de décomposition modulaire de G, au dessus on a indiqué le graphe premier correspondant au noeud P.

Le but recherché est d'obtenir une décomposition qui nous permette - malgré un graphe qui peut être complexe - de déduire un maximum de relations entre les sommets à partir de l'arbre de décomposition modulaire. Par exemple sur la figure 1.4c), on déduit grâce à cet arbre des liens qui existent entre les sommets 2, 3 et 4 ou 8, 9, 10 et 11.

Remarque 2 *Si l'on ne regarde que l'arbre de décomposition modulaire, le rôle de 1 et de 5 n'est pas parfaitement défini, tels qu'ils sont représentés leurs rôles sont équivalents. Comme tous les enfants du noeud P ont un rôle semblable, il est intéressant de pouvoir connaître la nature des liens qui les unissent.*

[2]Cette explication peut être trouvé dans le document à l'adresse : http ://philippe.gambette.free.fr/SCOL/Graphes.pdf.

1.3 Modules et réseaux métaboliques

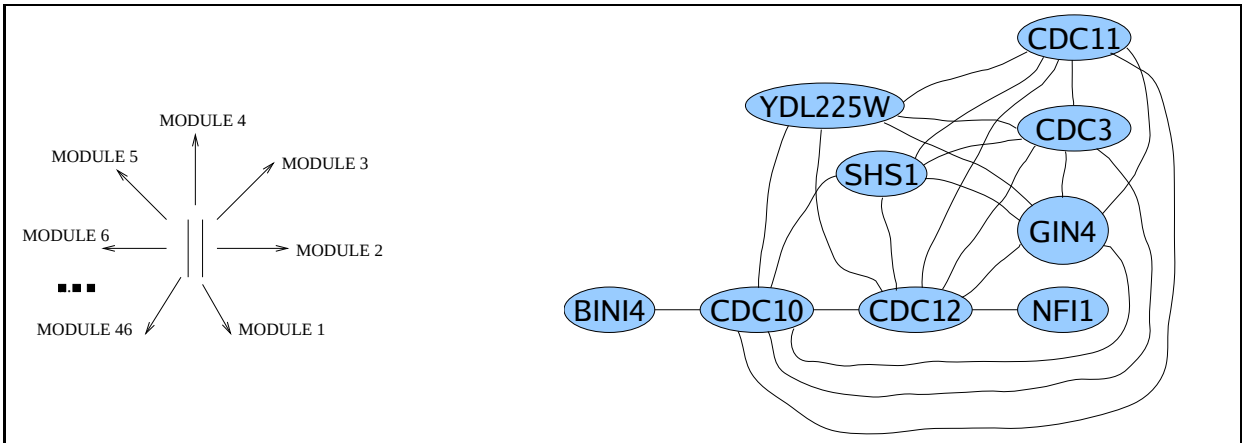


FIG. 1.5: La décomposition modulaire et les réseaux biologiques : application à la levure. A gauche, (a) la structure de la décomposition modulaire, un noeud parallèle composé de 46 modules fils. A droite, (b) Réseau associé à un module de la décomposition modulaire, nous avons choisi de nous focaliser sur un seul module car le graphe entier est beaucoup trop grand. (Il faut remplacer sur le schéma le nom de la protéine *BINI4* par *BNI4*)

La figure 1.6 représente le module de la figure 1.5b) issu de la décomposition modulaire du réseau d'interactions de protéines de la levure (figure 1.5a). [Yeast Interactome, Boston University, http ://structure.bu.edu/rakesh/myindex.html]

Remarque 3 *Les résultats généraux sur les modules séries et parallèles que nous allons présenter ont été déduit par [Gagneur et al., 2004]. Il a travaillé notamment sur l'exemple du réseau de complexes de régulation transcriptionnel présenté dans les résultats expérimentaux. Par soucis de clarté nous illustrerons la démarche générale sur l'exemple de la levure et non sur les résultats publiés dans [Gagneur et al., 2004].*

Les modules séries Dans un réseau d'interactions de protéines, [Gagneur et al., 2004] montre qu'un module série regroupe des protéines qui coopèrent pour accomplir une certaine fonction[3]. Sur la figure 1.7 qui représente un module série issu de la décomposition de la figure 1.6, toutes les protéines sont en relation les unes avec les autres.

Les modules parallèles Les protéines qui composent un module parallèle réalisent chacune la même (ou proche) fonction. Un exemple de telles protéines se trouve figure 1.6 avec les protéines $SHS1$ et $YDL225W$. On retrouve ces protéines figure 1.7 où le module parallèle qui les contient est inclus dans un module série.

[3]La fonction d'une protéine ou d'un groupe de protéines fait référence à une activité métabolique, le résultat d'une activité biologique, et des structures cellulaires ou sub-cellulaires. (Une définition peut être trouvée à l'adresse : http ://bioinfo.unice.fr/enseignements/www2005/documentation/Les%20enjeux%20de%20la%20biologie%20virtuelle.htm)

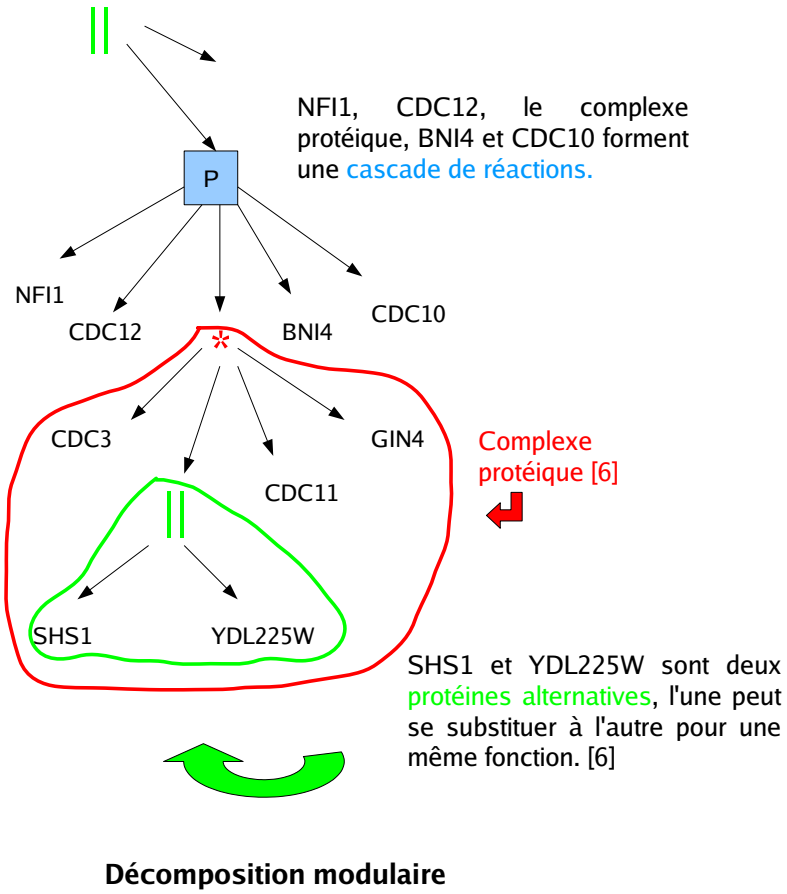


FIG. 1.6: La décomposition modulaire et les réseaux biologiques : application à la levure. Les protéines $CDC3$, $CDC11$, $GIN4$ et le complexe composé de $SHS1$ et $YDL225W$, forment un complexe protéique où chaque protéine est essentielle pour assurer la fonction globale du complexe.

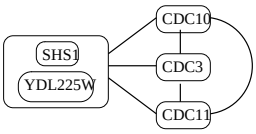


FIG. 1.7: Le module série issu de la décomposition figure 1.6.

Les noeuds premiers (N ou P) Tous les fils de ce type de noeud forment une cascade de réactions, mais on ne peut pas déterminer exactement les liens qui existent entre ses fils. Sur la figure 1.8, $BNI4$ et $CDC10$ sont connectés. On a pu effectuer cette connection seulement grâce à la figure 1.5 b) qui montre que $BNI4$ a pour seule connection $CDC10$. L'arbre de la figure 1.6 ne peut pas prévoir cette connection. On veut pouvoir mieux cerner le rôle de chaque protéine indépendamment et également au sein de l'éventuel module auquel elle appartient, et l'un des objectifs majeur en biologie est de pouvoir discerner le rôle de chaque protéine.

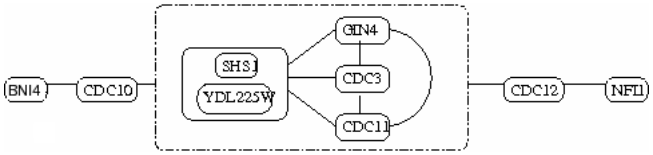


FIG. 1.8: Module P issu de la décomposition figure 1.6. On peut voir la suite des réactions entre les protéines. Le module série figure 1.7 n'est plus qu'une brique de la cascade de réactions.

1.4 Limite de la décomposition modulaire

La Remarque 2 est renforcée par l'existence de graphes comme celui de la figure 1.9. La décomposition modulaire ne décrit pas de liens entre les sommets du graphe.

FIG. 1.9: Limite de la décomposition modulaire a) le graphe G b) l'arbre de décomposition modulaire de G. On ne peut déduire de cet arbre les relations qui peuvent exister entre les fils du noeud P.

Chapitre 2

La décomposition homogène : aspects théoriques

La décomposition homogène permet de mieux comprendre les relations qui unissent les éléments d'un noeud de type premier. Cette décomposition étend la décomposition modulaire dans le sens où elle complète les informations données par la décomposition modulaire.

2.1 Définitions

La $p-$connexité ([Babel and Olariu, 1999] et [Babel and Olariu, 1997]) généralise la connexité et mène vers une représentation arborescente unique des graphes quelconques.

Définition 4 ***Graphe p-connecté***
Un graphe induit C de G est dit ***p-connecté*** *si pour toute partition de l'ensemble des sommets de C en deux ensembles non vides C_1 et C_2 de ses sommets, il existe un P_4 de G contenant des sommets à la fois de C_1 et de C_2. (Ce P_4 est appelé P_4* ***traversant****))*

Définition 5 ***Graphe caractéristique ou graphe quotient***
Le graphe obtenu depuis un graphe p-connecté G en contractant chaque ensemble homogène maximal en un seul sommet est appelé le ***graphe caractéristique ou graphe quotient*** *de G.*

Un exemple de tel graphe est présenté dans la figure 2.1b).

FIG. 2.1: Graphe caractéristique d'un graphe G a) Graphe G. b) Graphe caractéristique de G.

L'intérêt du graphe caractéristique réside dans le fait que lorsqu'on se trouve en présence d'un noeud de type P, on ne connaît pas à priori la forme du graphe (relations éventuelles entre ses fils), le graphe caractéristique donne la forme générale du graphe.

Définition 6 *$p-$ composante* ***ou composante*** *$p-$**connectée)***
Un sous graphe induit C p-connecté maximal est appelé ***p-composante****.*

Proposition 2 *Tout graphe possède une partition en $p-$ composantes.*

C'est la partition en $p-$ *composantes* d'un graphe (dont l'existence est assurée par la proposition 2) que nous allons récupérer lors de la décomposition homogène.

Il existe des $p-$ *composantes* qui sont séparables en deux ensembles disjoints comme décrit par la définition 7. Avec cette définition, la décomposition homogène s'appuie sur le théorème 3 pour décomposer au mieux les $p-$ *composantes* qui peuvent l'être.

Définition 7 *$p-$ composante* ***séparable***
Si un graphe C est une $p-$ composante et si ses sommets peuvent être partitionnés en deux ensembles disjoints C_1 et C_2 tels que tout P_4 traversant a ses points milieux dans C_1 et ses points finaux dans C_2 alors C est dit $p-$ composante séparable.
Par ailleurs, (C_1, C_2) est une séparation de G.

Un exemple de $p-$ *composante* séparable est présenté dans la figure 2.2.

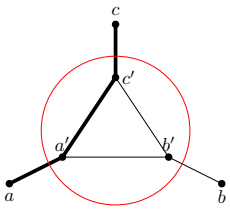


FIG. 2.2: Cette $p-$ *composante* séparable peut être partitionnée en $C_2 = \{a', b', c'\}$ contenant les points milieux des P_4 *traversants* du graphe et $C_1 = \{a, b, c\}$ contenant les points finaux correspondants. Par exemple pour le P_4 *traversant* $a - a' - c' - c$ on a a', $c' \in C_2$ et a, $c \in C_1$.

2.2 Théorèmes et algorithme de décomposition

Théorème 3 *[Jamison and Olariu, 1995]*
Tout graphe p-connecté séparable a une unique séparation. Chaque sommet appartient à un P_4 traversant.

Théorème 4 *[Jamison and Olariu, 1995]*
Si G est p-connecté séparable avec la séparation (C_1, C_2), le sous graphe de G (respectivement $\overline{G}$) induit par C_2 (respectivement C_1) est non connexe. Et tout $p-$ composante du sous graphe de G (respectivement $\overline{G}$) induit par C_2 (respectivement C_1) avec au moins 2 sommets est un ensemble homogène de G.

Une définition est nécessaire afin de décrire une classe de graphes particulière qui sera utilisée de le corollaire suivant :

Définition 8 ***Split*** *Un* ***split*** *est un graphe dont l'ensemble des sommets se décompose en deux ensembles disjoints, une clique K où tous les éléments sont adjacents et un stable S où tous les éléments sont non adjacents*

Corollaire 5 *Un graphe p-connecté est séparable si et seulement si son graphe caractéristique est un split.*

Ce corollaire induit la limite de la décomposition homogène.

Le plus important résultat dérivant de la p-connexité est le théorème structurel 6. Il implique une unique représentation arborescente d'un graphe quelconque.

Théorème 6 *Théorème structurel [Jamison and Olariu, 1995]*

Soit un graphe $G = (V, E)$, $|V| \geq 2$ quelconque. Exactement une seule de ces propositions est satisfaite :

- *G est non connexe.*
- *$\overline{G}$ est non connexe.*
- *Il existe une p-composante séparable propre*[1] *H de G avec une partition (H_1, H_2) tel que tout sommet n'appartenant pas à H est adjacent à tous les sommets de H_1 et aucun de H_2.*
- *G est p-connecté.*

La décomposition qui découle naturellement de ce théorème est appelée décomposition primitive. Il s'agit de décomposer le graphe $G = (V, E)$ selon les opérations op_i avec $i \in \{0, 1, 2\}$.

- Si G est non connexe, l'opération op_0 consiste à créer un noeud de type 0 dont les enfants sont les composantes connexes de G. Cette opération correspond au noeud de type parallèle de la décomposition modulaire.
- $\overline{G}$ est non connexe, l'opération op_1 consiste à créer un noeud de type 1 dont les enfants sont les composantes connexes de $\overline{G}$. Cette opération correspond au noeud de type série de la décomposition modulaire.
- Il existe une p-composante séparable propre H de G avec une partition $H = (H_1, H_2)$ tel que tout sommet n'appartenant pas à H est adjacent à tous les sommets de H_1 et aucun de H_2. L'op_2 consiste à créer un noeud de type 2 dont les enfants sont les noeuds qui sont racines des sous arbres issus du rappel de l'algorithme sur le graphe induit par les éléments n'appartenant pas à H, et sur le graphe induit par V - les éléments n'appartenant pas à H.

Proposition 7 *[Jamison and Olariu, 1995]*

Tout graphe peut être obtenu uniquement à partir de ses $p-$composantes par une séquence finie des opérations op_0, op_1 et op_2 .

L'arbre issu de cette décomposition (arbre primitif) est composé de noeuds internes $i \in \{0, 1, 2\}$. Un noeud i indique que le graphe associé au sous arbre de racine i est obtenu à partir des graphes correspondants à ses enfants par une i-opération. Les feuilles sont les $p-$*composantes* de G.

Arbre de décomposition homogène Le quatrième point du théorème permet de déterminer deux autres types de noeuds faisant référence à deux opérations op_3 et op_4.

- Soient $G_0 = (V_0 \cup \{y_0, y_1, \cdots, y_t\}, E_0)$, $G_1 = (V_1, E_1)$, $\cdots$, $G_t = (V_t, E_t)$ des graphes quelconques. Le graphe $G = (V, E)$ provient de G_0 et $G_1, \cdots, G_t$ par une op_3 s'il correspond

[1] Dont l'ensemble des sommets n'est ni égale à l'ensemble vide ni à V

au graphe G_0 dans lequel tous les sommets $y_i,\ i\ \in \{0, \cdots, t\}$ ont été remplacés par le graphe G_i. L'op_3 consiste à créer un noeud de type 3 dont les fils sont ces G_i.

- Si la $p - composante$ caractéristique est un split alors on peut définir l'opération op_4 en plaçant la clique K en racine du noeud op_4, et pour tous les sommets s du stable S il existe une clique $s \cup N(s)$ (avec $N(s)$ voisinage du sommet s) que l'on place en enfant du noeud 4.

Ces opérations seront explicitées dans le chapitre suivant.

La décomposition homogène affine la décomposition primitive et permet de décomposer un graphe en un arbre dont les noeuds internes correspondent à l'ensemble des opérations citées ci-dessus (à savoir op_0, op_1, op_2, op_3 et op_4) par l'Algorithme 1 issu de [Jamison and Olariu, 1995]. Pour plus de précision sur les utilisations des différentes opérations, se référer à leurs descriptions dans les paragraphes précédents.

Cet algorithme n'est pas aisément implémentable en l'état. C'est pourquoi S. Baumann l'a transformé comme expliqué dans le chapitre suivant.

Algorithme 1 Construction de l'arbre de décomposition homogène de G par [Jamison and Olariu, 1995], Entrée : Un graphe quelconque $G = (V, E)$, Sortie : L'arbre de décomposition homogène correspondant à G.

si $|V| = 1$ **alors**
  Renvoyer l'arbre composé de l'unique sommet de G
sinon
  si G est p-connecté **alors**
    Soit $C(G)$ le graphe caractéristique de G. {Il s'agit du **quatrième point du théorème 6**, on fait appelle à l'op_4 et l'op_3.}
    On décompose $C(G)$ par une op_4.
    Soit α le noeud 4 racine d'un arbre T',
    Soient $Y_1, Y_2, \cdots, Y_t$ les ensembles homogènes maximaux de G.
    Soient $T_1, T_2, \cdots, T_p$ les arbres homogènes associés enracinés respectivement en $r_1, r_2, \cdots, r_p$. {On va décomposer G avec une op_3.}
    Renvoyer l'arbre obtenu en ajoutant $\alpha, r_1, r_2, \cdots, r_p$ comme enfants d'un noeud 3.
  sinon
    si G est non connexe **alors**
      Soient $G_1, G_2, \cdots, G_p$ $(p \geq 2)$ des composantes connexes de G.
      {**Premier point du théorème 6**, on fait appelle à l'op_0}
      Soient $T_1, T_2, \cdots T_p$ les arbres homogènes associés de racines respectives $r_1, r_2, \cdots, r_p$.
      Renvoyer l'arbre obtenu en ajoutant $r_1, r_2, \cdots, r_p$ comme enfants d'un noeud 0.
    sinon
      si $\overline{G}$ est non connexe **alors**
        {**Deuxième point du théorème 6**, on fait appelle à l'op_1.}
        Soient $\overline{G_1}, \overline{G_2}, \cdots, \overline{G_p}$ $(p \geq 2)$ des composantes connexes de $\overline{G}$.
        Soient $T_1, T_2, \cdots T_p$ les arbres homogènes associés de racines respectives $r_1, r_2, \cdots, r_p$.
        Renvoyer l'arbre obtenu en ajoutant $r_1, r_2, \cdots, r_p$ comme enfants d'un noeud 1.
      sinon
        {G **satisfait le point 3 du théorème 6**, on fait appelle à l'op_2.}
        Ecrire G sous la forme $G_1 op_2 G_2$.
        Soient T_1, T_2 les arbres homogènes correspondants de racines respectives r_1 et r_2.
        Renvoyer l'arbre obtenu en ajoutant r_1 et r_2 comme enfants d'un noeud 2.
      fin si
    fin si
  fin si
fin si

Chapitre 3

La décomposition homogène : calcul de l'arbre de décomposition

Le calcul pratique de l'arbre de décomposition décrit dans [Bauman, 1996] se base sur la décomposition modulaire.

3.1 Principe de l'algorithme développé par Stephan Baumann

Définition 9 *Graphe caractéristique (complément)*

Le ***graphe caractéristique*** *d'un noeud dans un arbre T issu de la décomposition d'un graphe $G = (V, E)$, correspond à un graphe $G' = (V', E')$ où V' est l'ensemble contenant un sommet de chaque fils de ce noeud et E' les arêtes existantes entre ces sommets dans G. L'ensemble V' est appelé* ***système de représentants****. Cette définition complète de façon plus intuitive la définition 5 du chapitre précédent.*

La décomposition homogène complète la décomposition modulaire en optimisant la description des noeuds de type P (voir la limite de la décomposition modulaire section 1.4). On récupère pour cela dans un premier temps tous les systèmes de représentants de chacun des noeuds P. Ce sont ces ensembles de sommets que l'on va traiter. Selon leurs propriétés on aboutit à un noeud de type différent dans l'arbre. Ces différents cas sont détaillés dans le paragraphe suivant ainsi que l'algorithme 2 correspondant.

Algorithme 2 Construction de l'arbre de décomposition homogène de G par [Bauman, 1996], Entrée : L'arbre de décomposition modulaire de $G = (V, E)$ obtenu avec l'algorithme de [McConnell and Spinrad, 1994], Sortie : L'arbre de décomposition homogène correspondant à G.

Calcul du système de représentants Y_β pour chaque noeud β de type P de la décomposition modulaire issue de [McConnell and Spinrad, 1994].
pour Tous les Y_β **faire**
  si Y_β est un split où K est une clique et S un stable **alors**
    si $Y_\beta = K \cup S$ et tous les éléments de K sont adjacents à au moins un élément de S **alors**
      {**Correspond au cas 1 section 3.2**}
      Renommer l'étiquette de β en NOEUD 3.
      Supprimer les fils de β qui sont des feuilles.
      CRÉER UN NOEUD 4 étiqueté avec K et replacer le tout comme enfant de β.
      Étiqueter les enfants de 4 avec $s \cup (N(s) \cap K)$ pour tous les $s \in S$ avec $N(s)$ voisinage de s.
    sinon
      $\{Y_\beta = R \cup \widetilde{K} \cup S$ où $\widetilde{K}$ contient tous les éléments de K adjacents à au moins un élément de S, et l(es) élément(s) de R sont adjacent(s) à tous les éléments de K et aucun de S.$\}$
      CRÉER UN NOEUD 2, SOIT α CE NOEUD.
      Renommer l'étiquette de β en NOEUD 3.
      Supprimer dans les fils de β, les feuilles et R.
      Parent(α) $\leftarrow$ Parent(β). {Si s est la racine alors Parent(s) = s par convention, sinon Parent(s) correspond au père direct du sommet s dans l'arbre.}
      Parent(β) $\leftarrow \alpha$.
      si $|R| = 1$ **alors**
        {R contient un sommet isolé, soit γ ce noeud. **Correspond au cas 2a section 3.2**}
        Parent(γ) $\leftarrow \alpha$.
      sinon
        {R représente un module représenté par un noeud γ. **Correspond au cas 2b section 3.2**}
        Parent(γ) $\leftarrow \alpha$.
        CRÉER UN NOEUD 4 étiqueté avec $\widetilde{K}$.
        Les enfants de $\widetilde{K}$ sont étiquetés avec $s \cup N(s) \cap \widetilde{K}$ pour tous les $s \in S$ avec $N(s)$ voisinage de s.
        Replacer le tout comme enfant de β.
      fin si
    fin si
  sinon
    {Ne rien faire.}
  fin si
fin pour

3.2 Illustration de différents cas de l'algorithme 2

Le corollaire 5 implique que les cas suivants n'apparaissent que si le graphe caractéristique est un split :

3.2.1 Cas 1 : création d'un noeud d'opération 4

Le graphe caractéristique $G' = (V', E')$ du graphe $G = (V, E)$ figure 3.1a), dont ici le système de représentants est égal à l'ensemble des sommets V, est un split avec $V' = K \cup S$ et tous les éléments de K sont adjacents à au moins un élément de S (figure 3.1)

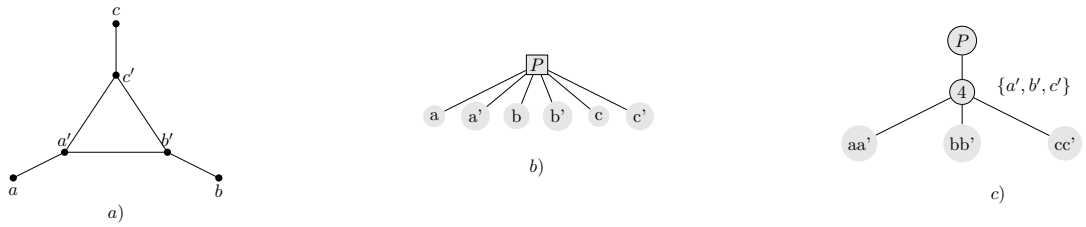


FIG. 3.1: Exemple de décomposition : cas d'un graphe $G = (V, E)$ où $V = K \cup S$ et tous les éléments de K sont adjacents à au moins un élément de S. a) Graphe G de départ. b) Arbre de décomposition modulaire de G. c) Arbre de décomposition homogène de G

Cette décomposition implique que le graphe se décompose en deux ensembles disjoints $K = \{a', b', c'\}$ la clique et $S = \{a, b, c\}$ le stable. Les étiquettes des feuilles du noeud 4 correspondent aux liens qui existent entre ce stable et cette clique. Par exemple on remarquera que le sommet $a \in S$ est couplé avec $a' \in K$ et que le graphe possède une arête entre a et a'.

3.2.2 Cas 2 : création d'un noeud d'opération 4 et un noeud d'opération 2

2a) Le graphe caractéristique $G' = (V', E')$ du graphe $G = (V, E)$ figure 3.2a), dont ici le système de représentants est égal à l'ensemble des sommets V, est un split avec $V' = \widetilde{K} \cup R \cup S$ et $|R| = 1$ (figure 3.2).

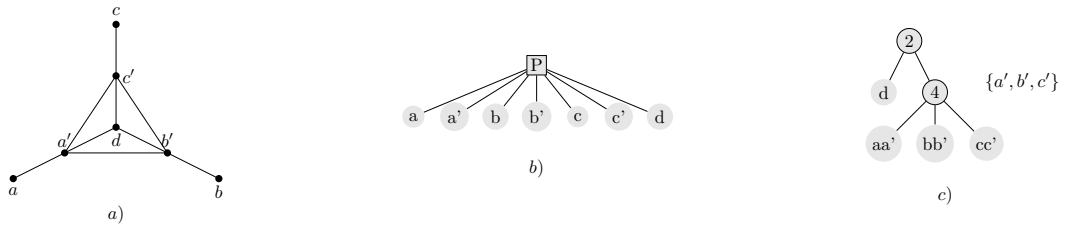


FIG. 3.2: Exemple de décomposition : cas d'un graphe $G = (V, E)$ où $V = \widetilde{K} \cup R \cup S$ et $|R| = 1$. Ici $R = \{d\}$, $\widetilde{K} = \{a', b', c'\}$ et $S = \{a, b, c\}$. a) Graphe G. b) Décomposition modulaire de G. c) Décomposition homogène de G.

Ici la décomposition est la suivante : $\widetilde{K} = \{a', b', c'\}$ et $K = \widetilde{K} \cup \{d\}$. $S = \{a, b, c\}$
Le sommet isolé d est relié à tous les éléments de la clique $\widetilde{K}$.

2b) Le graphe caractéristique $G' = (V', E')$ du graphe $G = (V, E)$ figure 3.3a) (le système de représentants associé est par exemple égal à $\{s_1, s_2, s_3, s_4, s_5, s_8, s_9\}$) est un split avec $V' = \widetilde{K} \cup R \cup S$ et R est un module γ comme le montre la figure 3.3.

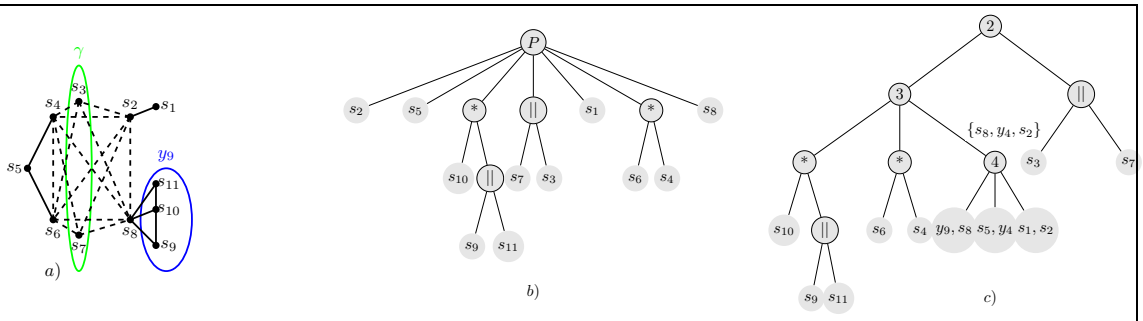


FIG. 3.3: Exemple de décomposition : cas d'un graphe $G = (V, E)$ où $V = \widetilde{K} \cup R \cup S$ et R est un module γ. Ici γ est le module parallèle entre les sommets s_3 et s_7, la clique $K = \{y_4, \gamma, s_2, s_8\}$ en pointillée et le stable $S = \{s_1, s_5, y_9\}$. a) Graphe G. b) Décomposition modulaire de G. c) Décomposition homogène de G.

La décomposition est la suivante : $K = \{y_4, s_2, s_8\} \cup \{s_3, s_7\}$ et $S = \{s_1, s_5, y_9\}$ et ici R est un module, en l'occurence il comprend les sommets s_3 et s_7. Ces deux sommets sont reliés à tous les éléments qui composent la clique du noeud 4 : $\{s_8, y_4, s_2\}$. Noter qu'ils sont également adjacents aux éléments qui composent y_4, y_4 représente le module fils du noeud P dont le sommet s_4 est un des descendants, tout comme y_9 que l'on peut identifier sur la figure 3.3.

Chapitre 4

A propos de l'implémentation

4.1 Implémentation de la décomposition modulaire par Julien Gagneur

[Gagneur et al., 2004] s'est appuyé sur l'algorithme de [McConnell and Spinrad, 1994] pour implémenter la décomposition modulaire. L'implémentation est en Java et la structure de donnée utilisée pour représenter un graphe $G = (V, E)$ est sous forme de liste d'adjacence. On peut obtenir, selon les choix de représentation, différents arbres de décomposition modulaire équivalents (par exemple figure 4.1 la représentation selon [Gagneur et al., 2004] en a) et celle selon [Bauman, 1996] en b), il est alors possible, avant de lancer la décomposition homogène, de transformer la décomposition modulaire en celle souhaitée ici par [Bauman, 1996] pour effectuer sa décomposition homogène.

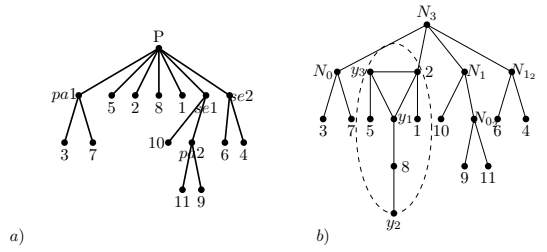


FIG. 4.1: a) Décomposition modulaire selon [Gagneur et al., 2004] b) Décomposition modulaire selon [Bauman, 1996]

4.2 L'implémentation de l'algorithme de Stephan Baumann

L'implémentation réalisée s'appuie sur la décomposition modulaire de [Gagneur et al., 2004], implémentée en java. L'algorithme est issu du rapport technique de [Bauman, 1996].

Les classes et fonctions principales implémentées Une classe (`NewVertex.java`) permet de rapprocher la structure de Julien Gagneur à celle de Stephan Baumann.

Il y a deux classes essentielles et la première (`edge_lists.java`) permet de passer de la représentation de Julien Gagneur à celle de Stephan Baumann (figure 4.1) en calculant pour chaque noeud P de l'arbre de décomposition modulaire le système de représentants associé. Une seconde classe (`DecompositionH.java`) effectue la décomposition proprement dite.

Il a aussi été nécessaire de compléter une classe de l'implémentation de Julien Gagneur (`ModuleTree.java`) afin de manipuler la structure d'arbre comme on le souhaitait. Pour de plus amples détails sur la programmation, se référer au code commenté en annexe[1].

4.2.1 L'algorithme principal de la classe edge_lists.java

Cet algorithme 3 passe de la structure d'implémentation de Julien Gagneur à celle de Stephan Baumann. Il récupère pour chaque noeud P de l'arbre de décomposition modulaire calculé par le programme de Julien Gagneur, la liste des sommets formant son système de représentants (voir définition 9).

Les notations utilisées Dans cet algorithme plusieurs notations ont été utilisées :

- s_α : le plus petit numéro d'un noeud de l'arbre enraciné dans un fils de α.
- c_α : nombre d'enfants du noeud α.
- Y_β : liste des sommets formant le système de représentants (voir définition 9) associé au noeud β de type premier.
- Les lettres minuscules tels que u, v correspondent à des noeuds, donc des positions dans l'arbre.

Remarque 4 *Dans l'algorithme 3, on a choisi de récupérer par convention le plus petit fils de chaque fils d'un noeud P. Dans la pratique on peut choisir n'importe lequel. L'algorithme aurait aussi sans doute été plus clair si un numéro de noeud et un noeud avaient été confondus. En fait, j'ai fait l'erreur de trop vouloir m'approcher de l'algorithme proposé dans [Bauman, 1996] en pensant qu'il serait plus propre que si j'avais moi même calculé les graphes caractéristiques, sans me rendre compte au premier abord qu'il n'était pas optimal.*

4.2.2 Les algorithmes nécessaires pour déterminer les différents cas (voir section 3.2) de la décomposition homogène

Ces algorithmes (implémentés dans la classe `DecompositionH.java`) utilisent le théorème 8 pour déterminer si un graphe est un split. Ce test et le corollaire 5 mettent en avant si le graphe est décomposable ou non par décomposition homogène.

Théorème 8 *Soit $G = (V, E)$, avec $|V| = n$, un graphe non orienté avec les degrés $d_1 \geq d_2 \geq \cdots \geq d_n$ et $m = \max\{i | d_i \geq i - 1\}$. G est un split si et seulement si $\sum_{i=1}^{m} d_i = m(m-1) + \sum_{i=m+1}^{n} d_i$.*

Remarque 5 *Par ailleurs, $d_m = m-1$ si et seulement si la partition de G en clique et en stable n'est pas unique. On se retrouve alors dans le cas 2 de l'algorithme de décomposition homogène.*

[1] Seules les classes edge_lists.java et DecompositionH.java sont présentes, pour obtenir l'intégralité du code vous pouvez écrire à geraldine.delmondo@gmail.com

Algorithme 3 Calcul du système de représentants de chaque noeud P de l'arbre de décomposition modulaire T_M de Julien Gagneur. Entrée : L'arbre de décomposition modulaire de $G = (V, E)$ obtenu avec l'implémentation de Julien Gagneur ainsi que le graphe G de départ sous la forme d'une liste d'adjacence de ses sommets. Sortie : Les graphes caractéristiques sous forme d'une liste de liste d'arêtes de chacune des listes des systèmes de représentants Y_β correspondant à chacun des noeuds β de type Premier.

Réduire la liste d'adjacence de G à une liste d'arêtes E telle que $E = \{ij | i < j\}$.
pour Chaque noeud α de T_M de type Série, Parallèle ou Premier **faire**
  $s_\alpha \leftarrow |V| + 1$ {On initialise à un grand nombre car **par convention** on prendra le plus petit enfant.}
  $d_\alpha \leftarrow 0$ {calcule le nombre d'enfants c_α du noeud α}
fin pour
{Initialisation à vide de la liste résultat qui sera composée des listes de représentants Y_β correspondantes à chacun des noeuds de type Premier de l'arbre T_M}
pour Chaque noeud β de T_M de type Premier **faire**
  Initialisation de la liste de représentants associée Y_β à vide.
fin pour
{On a considéré que l'ensemble des sommets V était de la forme $\{1, \cdots, n\}$.}
pour Toutes les feuilles $v \in \{1, \cdots, n\}$ **faire**
  $s_v \leftarrow v$; {On récupère le numéro de la feuille en cours.}
  $c_v \leftarrow 0$; $d_v \leftarrow 0$; {Une feuille n'a pas d'enfant}
  $u \leftarrow v$; {On sauve l'endroit où l'on se trouve dans l'arbre.}
  tant que $d_u = c_u$ et que u n'est pas la racine de T **faire**
    $d_{parent(u)} \leftarrow d_{parent(u)} + 1$ {Le parent de u à un enfant de plus (initialisation à 0).}
    {Par convention on a décidé de prendre le plus petit enfant, d'où la condition suivante :}
    si $s_u < s_{parent(u)}$ **alors**
      $s_{parent(u)} \leftarrow s_u$
    fin si
    {Si le parent de u est un noeud premier on doit sauver le numéro de son fils s_u comme représentant.}
    si $parents(u)$ est un noeud Premier **alors**
      $Y_{parent(u)} \leftarrow Y_{parent(u)} \cup \{s_u\}$
    fin si
    $u \leftarrow parent(u)$ {On continue en passant au parent de u}
  fin tant que
fin pour
{Comme on souhaite un graphe caractéristique, on récupère les arêtes associées à chaque Y_β.}
pour Toutes les arêtes $ij \in E$ **faire**
  si $j \in V_\beta - \{s_\beta\}$ et $i \in V_\beta$ **alors**
    Ajouter ij à la liste Y_β
  fin si
fin pour

L'algorithme 4 résume notre travail pour appliquer les résultats précédents à notre implémentation. Le détail de chacune des étapes est décrit en annexe dans le code commenté de la classe.

Algorithme 4 Entrée : Une liste d'arêtes correspondant au graphe sur lequel on veut pouvoir choisir l'un des cas de la décomposition homogène à appliquer (voir section 3.2). Sortie : Un couple de booléens (vrai,vrai) si E_β est décomposable par le cas 1, (vrai, faux) si c'est par le cas 2, (faux,faux) si ce n'est pas un split, on renvoie aussi la clique K et le stable S issus de E_β si possible

Calcul du degré de chacun des sommets apparaissant dans E_β.
Tri de ces degrés du plus grand au plus petit.
Calcul de m indice i maximal tel que $d_i \geq i - 1$.
Calcul de la Somme $\sum_{i=1}^{m} d_i$, au passage on récupère les sommets de la clique éventuelle.
Calcul de la Somme $\sum_{i=m+1}^{n} d_i$, au passage on récupère les sommets du stable éventuel.
Vérifications s'il s'agit d'un split à l'aide de la remarque 5.

Chapitre 5

Résultats expérimentaux

5.1 Les résultats obtenus par la décomposition homogène

5.1.1 Application au réseau de complexes de régulation transcriptionnel

Nous avons traité plusieurs des résultats issus de [Gagneur et al., 2004]. Nous nous intéressons ici à un réseau de 50 protéines définissant la restructuration de la chromatine par les complexes *RSC* et *SWI/SNF*, le complexe de facteurs de transcription général *TFIIF*, *TFIID*, et le complexe *MEDIATOR* qui gère les signaux pour l'activation de l'ARN polymérase II. Ces différents complexes ont été identifiés expérimentalement, et le résultat de ces expérimentations est illustré dans la figure 5.1. Le graphe associé à ces complexes est présenté dans la figure 5.2. Les résultats que nous obtenons pour ce réseau sont présentés figure 5.4 et 5.5. La visualisation de nos graphes est réalisée à l'aide du logiciel YED de chez yWorks[1]. Nous obtenons une décomposition modulaire similaire aux résultats publiés ainsi qu'une décomposition homogène plus précise dans la description de la hiérarchie entre modules (figure 5.5).

La décomposition homogène (figure 5.5) identifie tous les complexes protéiques de la figure 5.1, ce que ne permettait pas la décomposition modulaire de [Gagneur et al., 2004] (figure 5.3 et 5.4). [Gagneur et al., 2004] ne parvient notamment pas à situer la protéine *ANC1* dans les différents complexes. La décomposition homogène reconnait cette protéine comme adjacente aux protéines du module *#ARP9* composé des protéines *ARP9* et *ARP7*. On constate que ce module est adjacent au module fils du noeud 2 (par définition), ce qui reconstitue le complexe protéique *SWI/SNF*. De récentes études expérimentales [Szerlong et al., 2003] identifient les protéines du module *#ARP9* comme composants d'un sous-complexe qui faciliterait la re-structuration de la chromatine et les interactions entre complexes. Par ailleurs, le complexe *RSC* est également impliqué dans la restructuration chromatique et possède le sous complexe *ARP9* et *ARP7*. La décomposition homogène identifie un module *#STH1* couplé au module *#ARP9*, ce qui, ce qui confirme expérimentalement l'intérêt de notre démarche. La même analyse montre *ANC1* adjacente aux protéines du module *#TAF40*. On retrouve ainsi de manière fine les trois complexes protéiques (*MEDIATOR*, *TFIID* et *TFIIF*) mais avec une description supplémentaire qui peut être interprétée biologiquement.

[1] http ://yworks.com/en/index.html

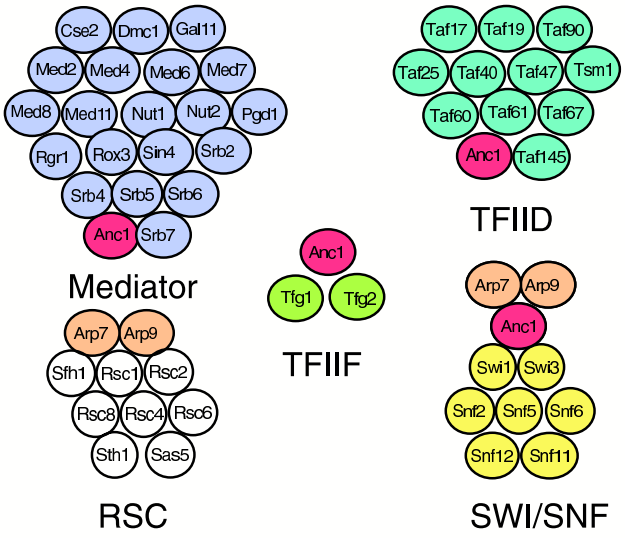


FIG. 5.1: La décomposition modulaire selon [Gagneur et al., 2004] (l'image est issue de l'article). Modélisation des complexes protéiques obtenus expérimentalement.

5.1.2 Application au réseau d'interactions de protéines de la levure

La figure 5.6 représente les décompositions modulaire et homogène du module (figure 1.5b) issu de la décomposition modulaire du réseau d'interactions de protéines de la levure. [Yeast Interactome, Boston University, http ://structure.bu.edu/rakesh/myindex.html][2]

Dans la décomposition modulaire, il n'était pas possible de connaître les liens entre les protéines *NFI1*, *CDC12*, *BNI4*, *CDC10* et le complexe protéique composé de *CDC3*, le module parallèle (*SHS1*, *YDL225W*, *CDC11* et *GIN4*. On complète sensiblement les informations fournies par la décomposition modulaire (figure 5.6).

5.2 La décomposition homogène ne complète pas toujours la décomposition modulaire

Malheureusement, plusieurs tests n'ont pas amélioré la décomposition modulaire de certains réseaux biologiques. Sur la base de données $KEGG$[3] notamment, les informations données par la décomposition modulaire des réseaux tels que le métabolisme de l'azote ou de la fixation des composés carbonés[4] n'ont pas pu être complétées par l'application de la décomposition homogène.

[2]Les figures 5.7 b) et 5.7 c) représentent respectivement la décomposition modulaire et la décomposition homogène tels que nous les avons obtenus avec notre implémentation et le logiciel Yed pour l'affichage.

[3]http ://www.genome.jp/kegg/

[4]Respectivement map00910 et map00710 sur http ://www.genome.jp/kegg/xml/map/index.html

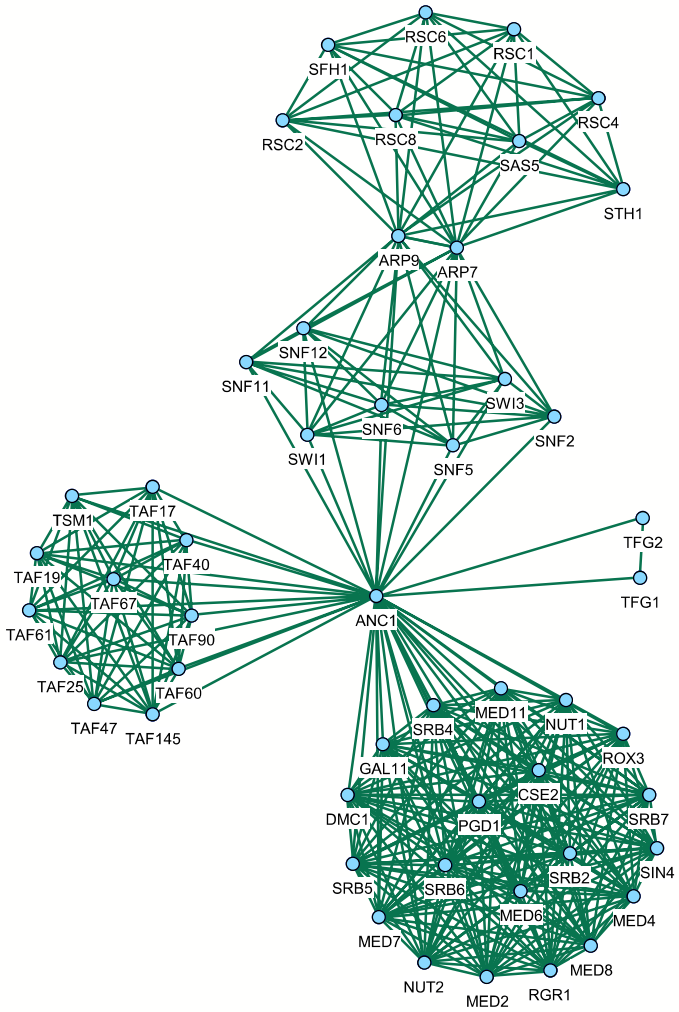


FIG. 5.2: La décomposition modulaire selon [Gagneur et al., 2004] (l'image est issue de l'article). Le graphe G associé aux complexes protéiques de la figure 5.1.

5.3 Implémentation de Roger Guimera et Luis A. Nunes Amaral

La comparaison de la décomposition homogène avec les travaux de [Guimera and Amaral, 2005] sur les deux tests que nous avons effectués (figure 5.8 et 5.9), donne des résultats différents. Les groupes de sommets issus de la décomposition homogène forment des ensembles plus petits. Par ailleurs, si notre décomposition donne les liens entre les groupes de sommets formés, ce n'est pas le cas de l'autre. La méthode de Roger Guimera et Luis A. Nunes donne une description en groupe de sommets mais ne décrit pas la hiérarchie entre ces groupes. La décomposition homogène peut être utile surtout si l'on souhaite des modules plus fins.

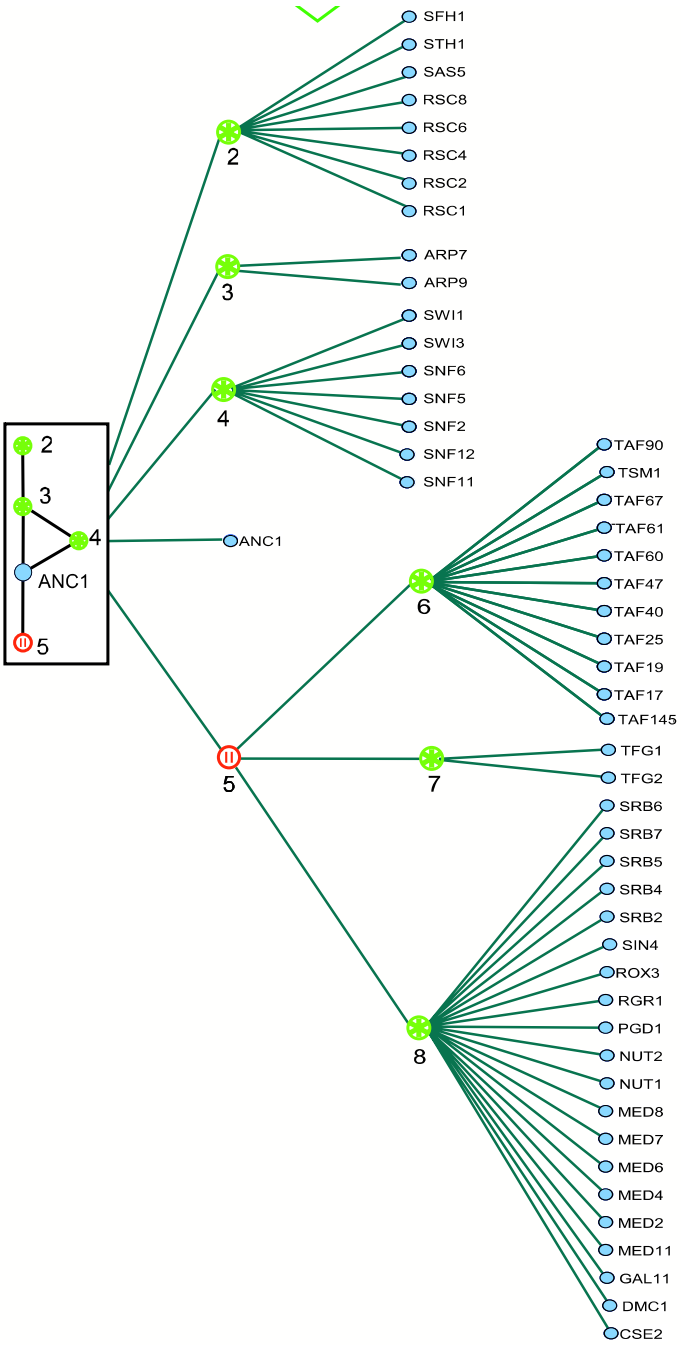


FIG. 5.3: La décomposition modulaire selon [Gagneur et al., 2004] (l'image est issue de l'article). Décomposition de G figure 5.2 par décomposition modulaire. Les noeuds 2, 3, 4, 6, 7 et 8 sont séries, le noeud 5 de type parallèle, la racine de type P.

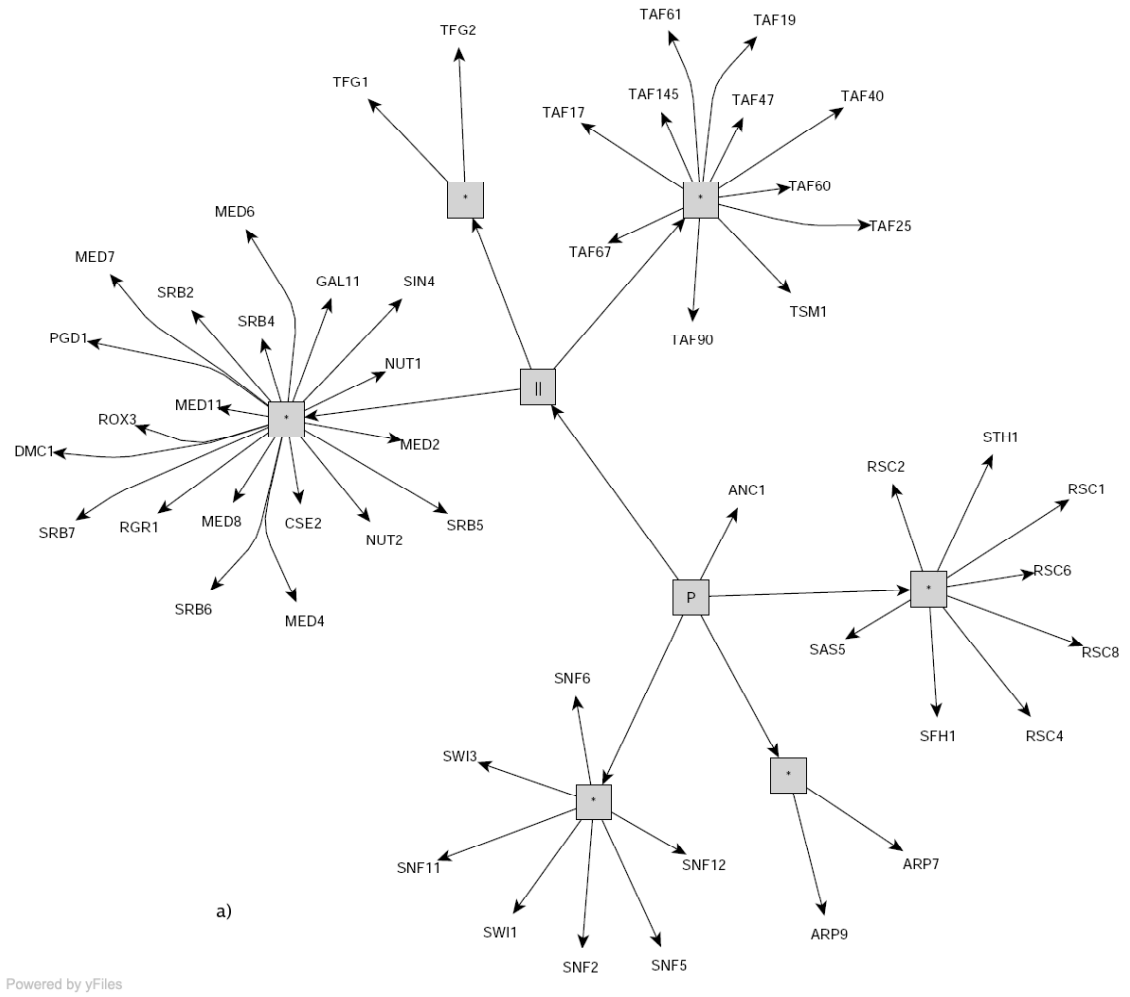


FIG. 5.4: La décomposition modulaire sur l'exemple de [Gagneur et al., 2004] : a) Décomposition modulaire du graphe figure 5.2, c'est le même résultat que la figure 5.3 mais tel que nous l'avons obtenu avec l'application de notre programme et l'affichage par le logiciel YED. Ce qui apparaît comme un point en sommet est en fait une * correspondant à un module série.

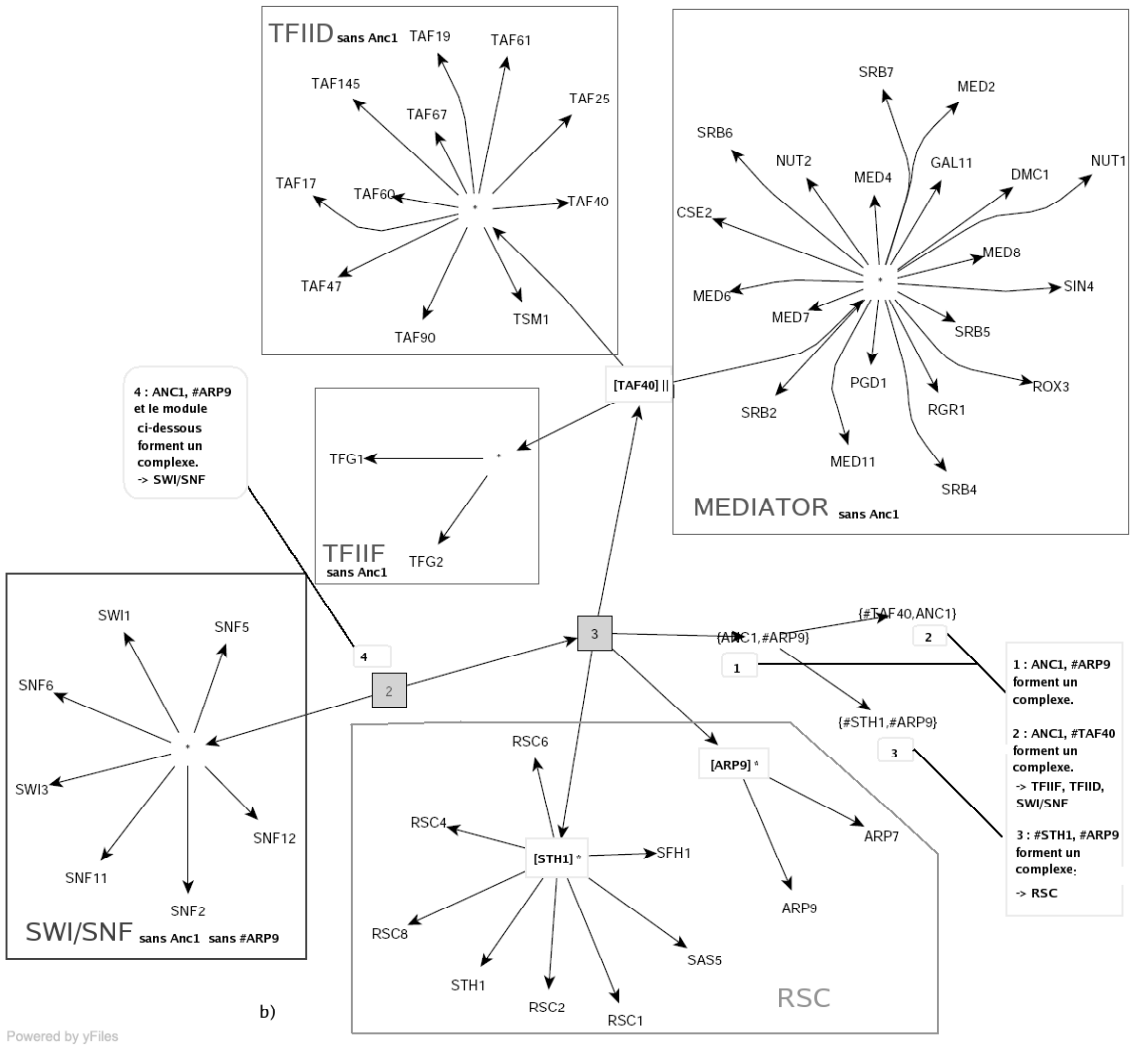


FIG. 5.5: Application de notre implémentation de la décomposition homogène sur l'exemple de [Gagneur et al., 2004] : Décomposition homogène du graphe figure 5.2. Ce qui apparaît comme un point en sommet est en fait une * correspondant à un module série. Dans cette décomposition homogène, certains noms de protéines sont précédés par le signe #, il s'agit dans ce cas de l'ensemble des protéines présentent dans le module auquel appartient la protéine marquée #, filles du noeud P correspondant. C'est pourquoi dans notre visualisation nous avons ajouté avec le type de module (* ou ||) la protéine marquée # lorsque c'était nécessaire pour retrouver plus facilement le module associé. Les numéros 1, 2, 3 et 4 ajouté à la figure explique les déductions que rend possibles la décomposition homogène, en particulier,le 1,2,3 sont liés à l'interprétation du noeud de type 4 et le 4 est lié à l'interprétation du noeud de type 2.

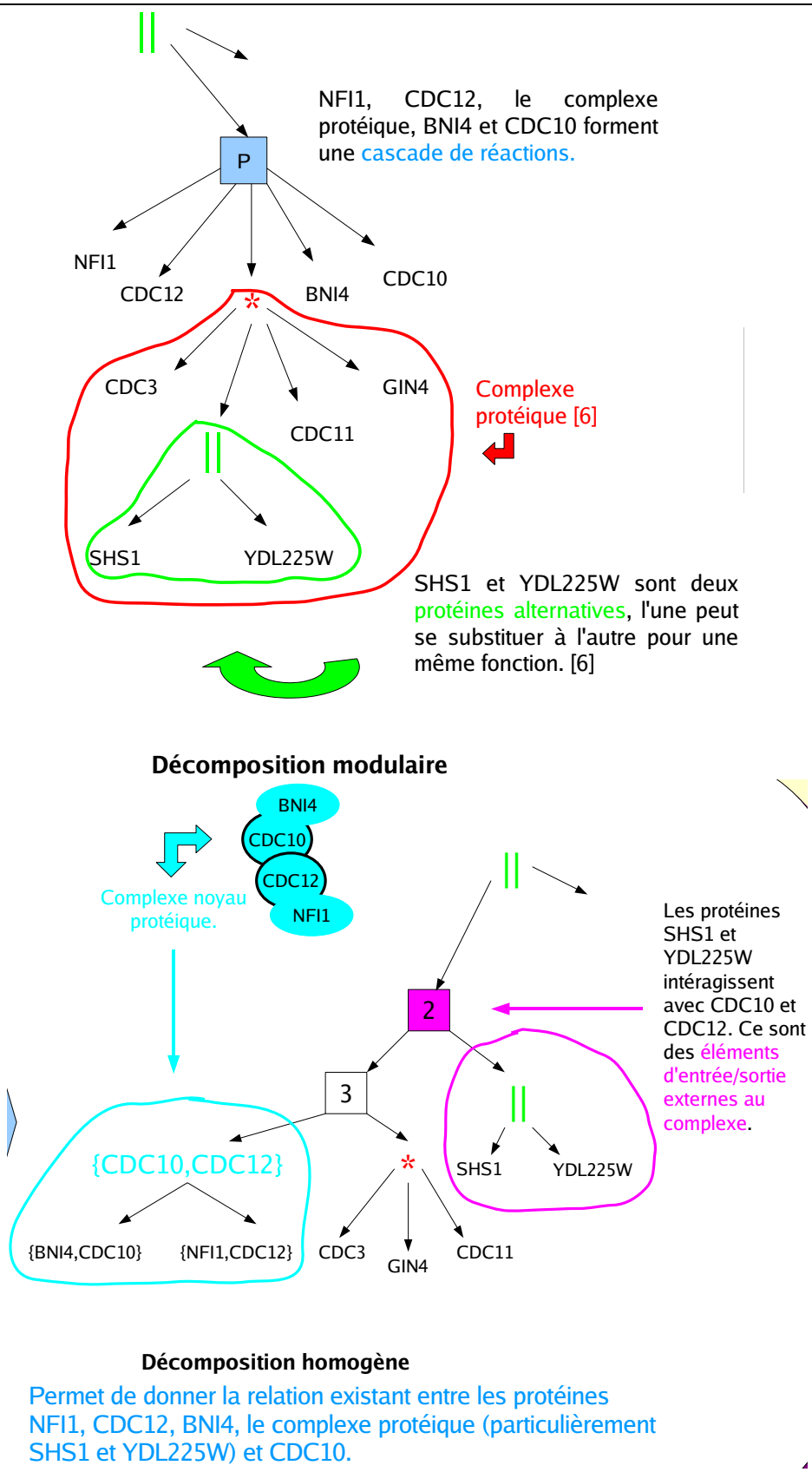


FIG. 5.6: La décomposition modulaire et la décomposition homogène dans les réseaux biologiques : application à la levure. L'arbre du bas complète celui obtenu par décomposition modulaire en haut, on peut en déduire qu'il existe un complexe noyau protéique composé du noyau de protéines *CDC10*, *CDC12*, et des protéines connexes *BNI4* et *NFI1*. Par ailleurs, les protéines *SHS1* et *YDL225W* sont des éléments d'entrée/sortie externes au complexe.

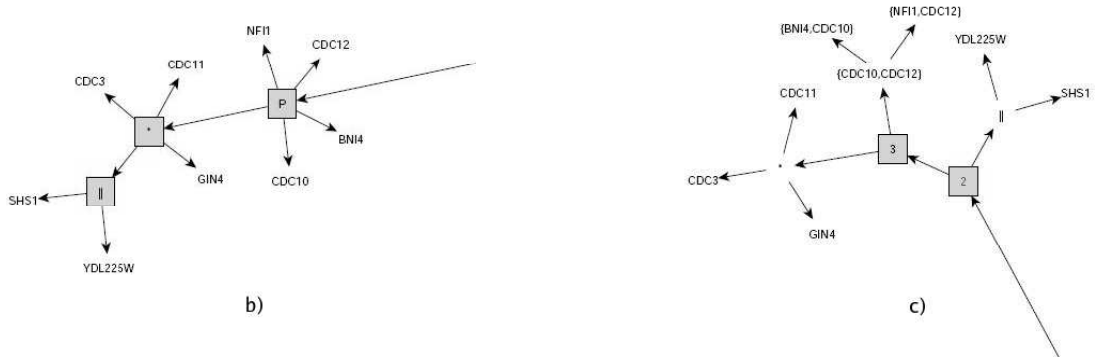


FIG. 5.7: Application de la décomposition modulaire et homogène au réseau d'interactions de la levure.(Complément de la figure 5.6) b) Vue détaillée de la décomposition modulaire du module. Les protéines *CDC3*, le module composé de *SHS1* et *YDL225W*, *CDC11*, *GIN4* forment un complexe protéique où chaque protéine est essentielle pour assurer la fonction globale du complexe. On ne connaît pas les liens entre ce complexe, *NFI1*, *CDC12*, *BNI4* et *CDC10*. c) Décomposition homogène du module. On peut en déduire qu'il existe un complexe noyau protéique composé du noyau de protéines $CDC10$, $CDC12$, et des protéines connexes *BNI4* et *NFI1*. Par ailleurs, les protéines *SHS1* et *YDL225W* sont des éléments d'entrée/sortie externes au complexe.

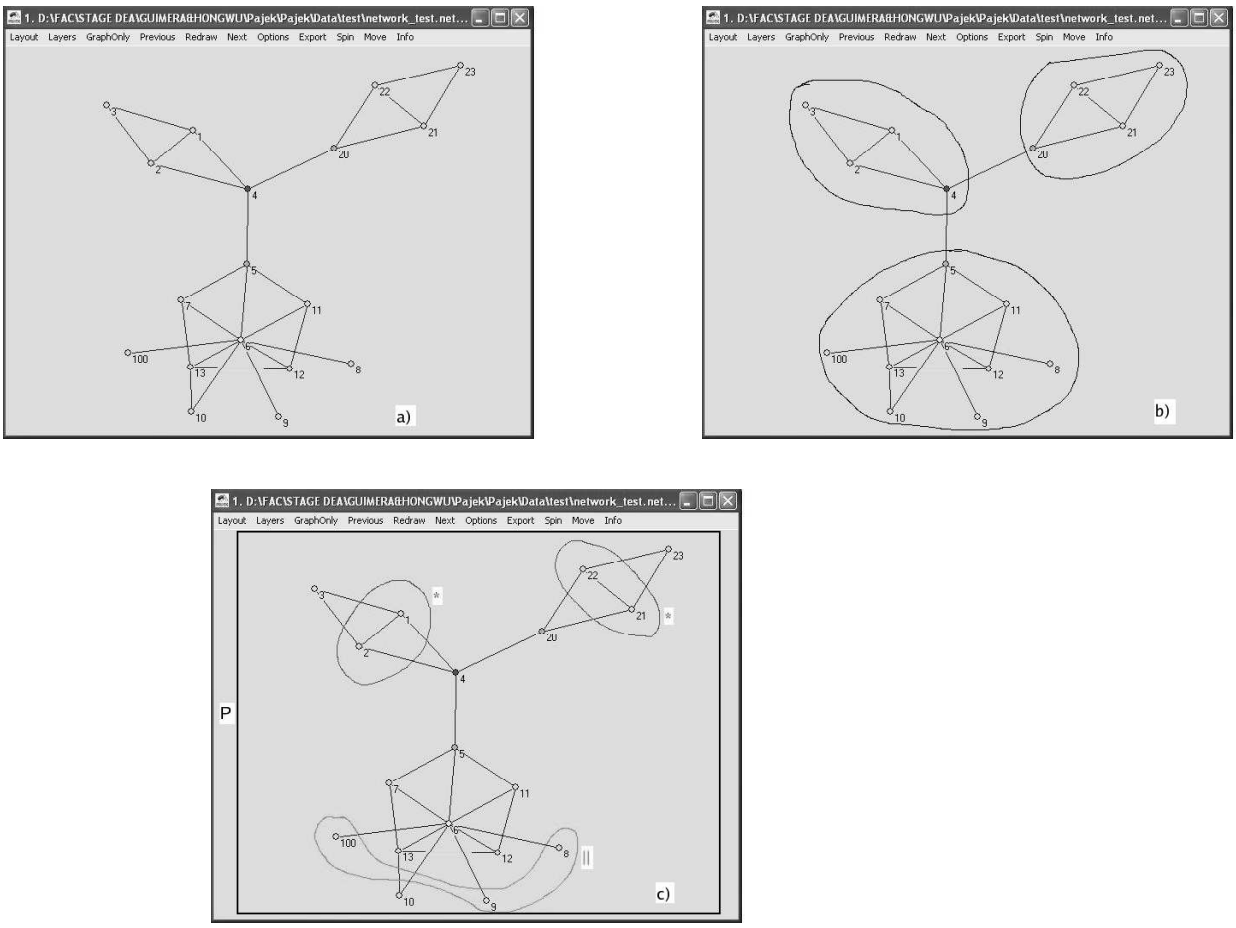


FIG. 5.8: Comparaison de la méthode de [Guimera and Amaral, 2005] avec la décomposition modulaire/homogène sur un graphe test fourni avec le programme de [Guimera and Amaral, 2005]. a) Graphe G de départ. b) Décomposition de G selon [Guimera and Amaral, 2005], on obtient trois groupes. c) Décomposition de G avec la décomposition modulaire (le résultat de la décomposition homogène est équivalent)

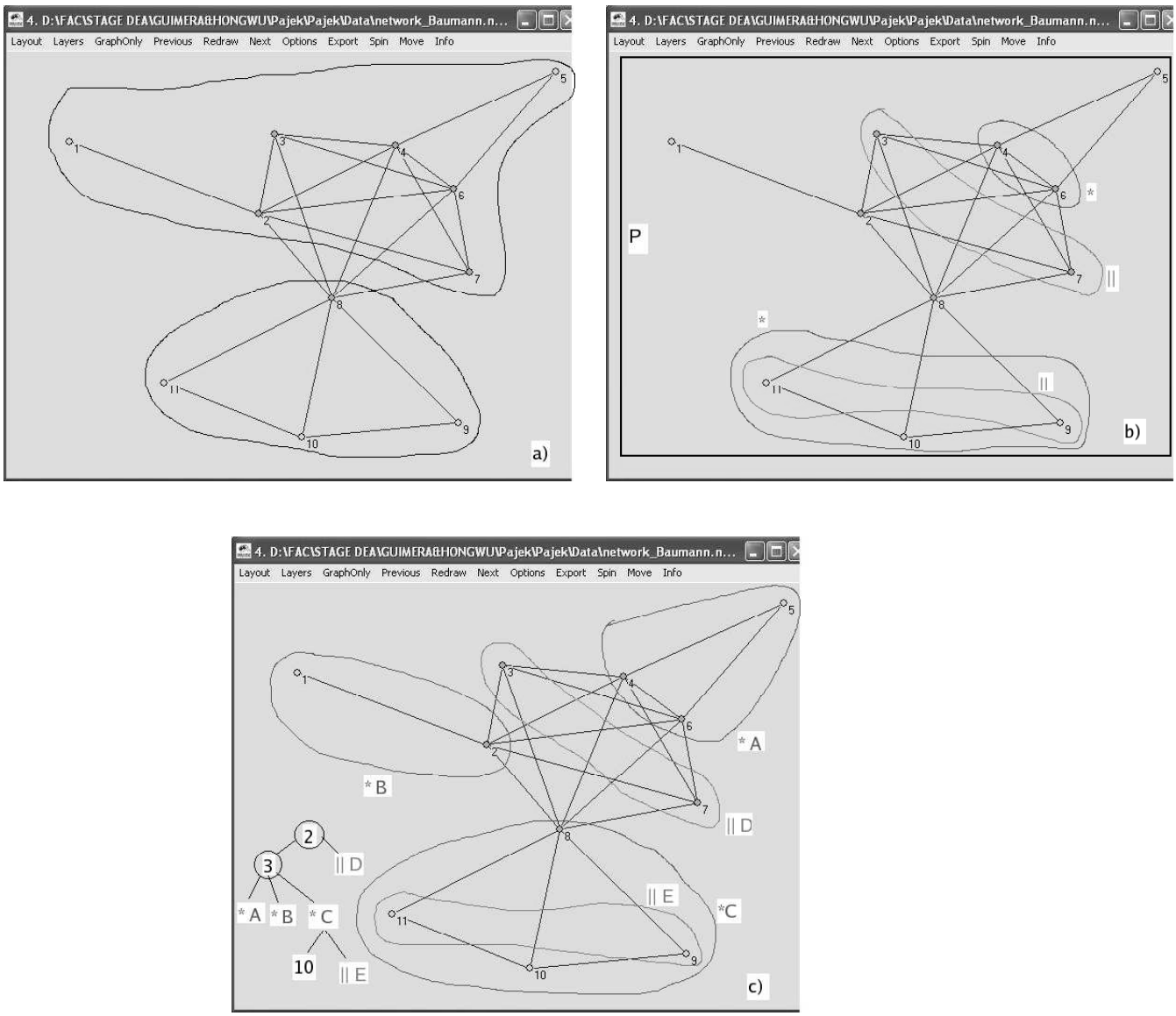


FIG. 5.9: Comparaison de la méthode de [Guimera and Amaral, 2005] avec la décomposition modulaire/homogène sur le graphe figure 3.3a. a) Méthode de [Guimera and Amaral, 2005] appliquée au graphe G de départ, on obtient deux groupes. b) Décomposition modulaire de G, l'arbre de cette décomposition se trouve en figure 3.3b. c) Décomposition homogène de G, l'arbre de cette décomposition se trouve en figure 3.3c).

Chapitre 6

Conclusion

6.1 Limite de la décomposition homogène

Il existe des classes de graphes pour lesquels la décomposition homogène n'apporte aucune information supplémentaire à la décomposition modulaire. Ce sont des graphes pour lesquels le graphe caractéristique n'est pas un split. Le noeud P n'est pas décomposé davantage. Le graphe de la figure 6.1 qui est celui de la figure 1.4 auquel on a ôté l'arête $\{7, 9\}$ en est un exemple caractéristique. L'arbre de décomposition homogène obtenu est le même que celui issu de la décomposition modulaire. Une intuition pour comprendre ce phénomène est qu'à partir du moment où le système de représentants comprend un nombre de noeuds trop important[1] la probabilité de trouver un split est faible car les contraintes qu'il entraîne sont fortes.

FIG. 6.1: Limite de la décomposition homogène a) le graphe induit par un système de représentants qui doit être un split pour que la décomposition homogène soit efficace. b) l'arbre de décomposition modulaire qui est aussi l'arbre de décomposition homogène.

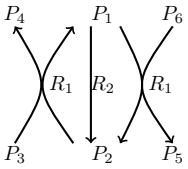
6.2 Formalisation du réseau

6.2.1 Le problème du traitement des arêtes multiples dans les réseaux biologiques

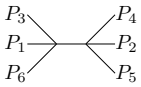
En fonction de la formalisation du réseau macromoléculaire, les graphes biologiques peuvent présenter des arêtes multiples. Elles correspondent à certains schéma de réactions biologiques particuliers comme le montre la figure 6.2. La question de leur modélisation est un point essentiel

[1] ce qui dépend évidemment du nombre de fils du noeud N, expérimentalement au delà de 5 on a peu de résultats

à résoudre pour que les résultats obtenus sur les graphes modélisant les réseaux biologiques soient le plus proches de la réalité possible.



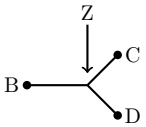
Le reseau metabolique
(1) $P_3 + P_1 + P_6 \rightarrow_{R_1} P_4 + P_2 + P_5$
(2) $P_1 \rightarrow_{R_2} P_2$



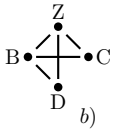
Graphe associe au reseau metabolique
Seule la relation (1) *est representee*

FIG. 6.2: Une arête multiple telle quelle est modélisée dans un réseau biologique. Les protéines sont étiquetés P et les réactions R. A gauche le réseau métabolique et les relations qu'il représente, à droite la modélisation d'une de ces relations sous forme de graphe.

Une modélisation possible Une méthode de décomposition doit être capable d'analyser un tel graphe. Nous proposons une modélisation de ce type d'arêtes comme illustré figure 6.3. Nous considérons qu'un substrat B catalysé par une enzyme Z et donnant les produits C et D (figure 6.3 a) est lié avec Z, C et D, et que Z est lié aux produits C et D. Par contre, les produits n'ont aucune raison à priori d'être liés entre eux, d'où la modélisation figure 6.3 b).



a)



b)

FIG. 6.3: Possibilité de modélisation d'une arête multiple, on représente les interactions de l'enzyme Z par les liens entre Z et les protéines concernées sur le graphe de modélisation. a) L'arête multiple, la protéine B catalysée par l'enzyme Z donne les produits C et D. b) Modélisation de l'arête multiple.

Intérêt de la modélisation dans un réseau Si l'on prend l'arête multiple "seule", la décomposition homogène n'apporte aucune information supplémentaire par rapport à la décomposition modulaire. Connectée à un élément, la décomposition homogène est plus précise que la décomposition modulaire (figure 6.4). Si la décomposition modulaire ne discerne pas les liens entre C, D, E et le module série (B,Z), la décomposition homogène apporte plus d'information en isolant le module série (B,Z) lié à C et à D, C est lié à E, C et D sont disjoints.

6.2.2 Intérêt de l'orientation des graphes ?

Par définition, une majorité des réseaux biologiques sont orientés. Une décomposition homogène de graphes orientés représente alors une perspective théorique intéressante. Dans ce cas, les modules obtenus lors de la décomposition d'un graphe orienté seraient un sous ensemble des modules obtenus lors d'une décomposition sur ce même graphe non orienté car les orientations

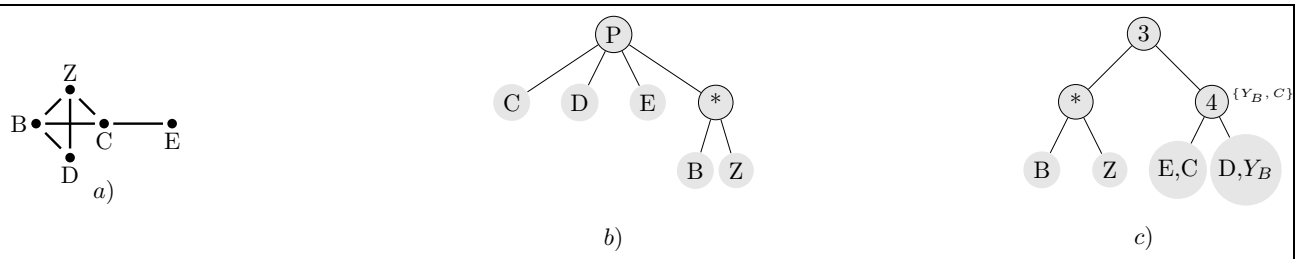


FIG. 6.4: Intérêt de la décomposition homogène dans le cas d'arêtes multiples. a) L'arête multiple (modélisée), la protéine B catalysée par l'enzyme Z donne les produits C et D, et le substrat C donne le produit E. b) Décomposition modulaire de l'arête multiple. On ne connaît pas les liens entre C, D, E et le module série (B,Z) c) Décomposition homogène de l'arête multiple. On en déduit que le module série (B,Z) est lié à C et D, C est lié à E, C et D sont disjoints

posent plus de contraintes. Par exemple, la clique à trois sommets non orientée nous donnera un module série, alors qu'en orienté il sera possible de trouver une orientation qui nous donne deux modules sur ce même graphe.

On notera qu'il existe une décomposition modulaire dans les graphes orientés, on peut en trouver une brève description dans l'habilitation à diriger la recherche de Christophe Paul [Paul, 2006]. Cependant il n'existe pas à ma connaissance de généralisation de la décomposition homogène aux graphes orientés. Ce problème biologique pourrait donc motiver des travaux théoriques fondamentaux en informatique.

6.3 Limite des décompositions dans les réseaux biologiques

L'importance de la modélisation en biologie La modélisation d'un système biologique (figure 6.5) [Kell and Knowles, 2006] passe par une perte d'information sur ce système. On choisira la formalisation du système en fonction de la question posée. Cela implique de garder à l'esprit qu'une modélisation du système n'est valide que dans un certain contexte et sous certaines conditions. Un modèle est donc une représentation limitée d'un système. Le choix de modélisation des réseaux de complexes protéiques influe énormément sur le résultat de la décomposition qui s'en suit. Ceci est illustré par les par les travaux [Wang and Zhang, 2007] qui décrivent deux types de modélisation souvent usitées. Pour le réseau de la figure 6.6a) si l'on choisit la modélisation "matrix"[2] de la figure 6.6b), la décomposition aura plutôt tendance à nous renvoyer des modules séries alors que la modélisation "spoke" de la figure 6.6c) nous renverra plutôt des modules parallèles. Suite à ces travaux, comment interpréter les différents types de modules que nous avons pu mettre en évidence ? Une voie de recherche est de mieux définir un module. Ce travail nécessite une collaboration étroite avec des biologistes qui peut aboutir à une nouvelle définition formelle, source de nouveaux algorithmes et recherches informatiques.

[2]Notez que Julien Gagneur utilise plutôt la modélisation "matrix"

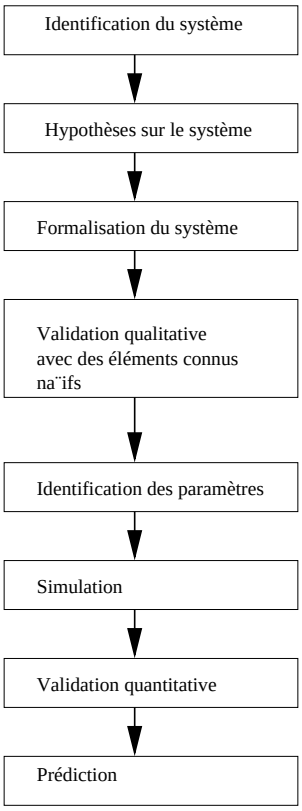


FIG. 6.5: Les étapes de la modélisation d'un système en biologie

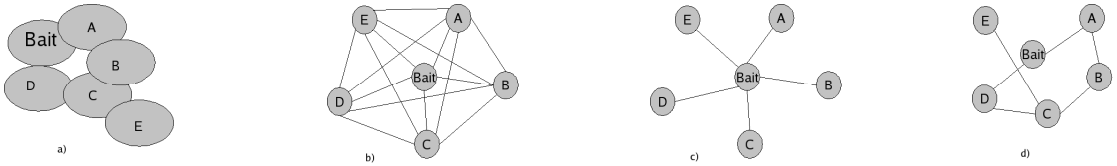


FIG. 6.6: Limite des décompositions sur les réseaux biologiques a) Complexe protéique. b) Modélisation "matrix" du complexe protéique. c) Modélisation "spoke" du complexe protéique. d) Modélisation réelle du complexe protéique.

6.4 Conclusion

Notre travail de décomposition des réseaux macromoléculaires s'est appuyé sur des outils théoriques de théorie des graphes que sont la décomposition modulaire et la décomposition homogène.

La décomposition modulaire implémentée par [Gagneur et al., 2004] permet de décomposer le réseau en modules et de connaître l'agencement des protéines. Il dénote trois types de modules, les modules séries, les modules parallèles et les modules premiers. S'il parvient à déduire des informations intéressantes des deux premiers types, le troisième reste beaucoup plus imprécis. Ce type de noeud étant malheureusement le plus répandu en biologie, il nous fallait donc résoudre

ce problème.

La décomposition homogène ([Jamison and Olariu, 1995], [Bauman, 1996]) s'appuie et étend la décomposition modulaire. Elle permet de décomposer plus précisément les modules de type premiers de la décomposition modulaire. Après implémentation, la décomposition homogène a été testée sur plusieurs réseaux biologiques dont un réseau de référence [Gagneur et al., 2004] et un réseau de la levure[3], la décomposition homogène complète dans les deux cas de manière significative les informations données par la décomposition modulaire.

Nous avons aussi testé notre programme sur des réseaux qui n'ont pas abouti sur des décompositions homogènes satisfaisantes. Les modules de types premiers n'ont pas été décomposés davantage.

Pour améliorer nos résultats il existe plusieurs voies de recherche connexes. L'une consiste à travailler sur des graphes orientés pour se rapprocher de la réalité biologique. Il faudrait pour cela développer un algorithme de décomposition homogène sur ces graphes. Une deuxième voie part d'un constat observé pendant nos expériences, des modules ne sont pas identifiés parce que la notion de module sur laquelle nous nous appuyons est trop rigide par rapport à la réalité biologique. Il s'agit de définir une notion de module plus lâche pour refléter la réalité biologique comme déjà initié par [Guimera and Amaral, 2005]. Les deux approches nécessitent une collaboration étroite avec les biologistes pour développer de nouvelles méthodes informatiques.

[3]Yeast Interactome, Boston University, http ://structure.bu.edu/rakesh/myindex.html

Bibliographie

[Adamcsek et al., 2006] Adamcsek, B., Palla, G., Farkas, I. J., Derenyi, I., and Vicsek, T. (2006). Cfinder : Locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22 :1021–1023.

[Babel and Olariu, 1997] Babel, L. and Olariu, S. (1997). On the separable-homogeneous decomposition of graphs (extended abstract). In *Proceedings of the 23rd International Workshop on Graph-Theoric Concepts in Computer Science*, volume 1335, pages 25–37.

[Babel and Olariu, 1999] Babel, L. and Olariu, S. (1999). On the p-connectedness of graphs - a survey. *Discrete Appl. Math.*, 95(1-3) :11–33.

[Barabási and Oltvai, 2004] Barabási, A. L. and Oltvai, Z. N. (2004). Network biology : understanding the cell's functional organization. *Nat Rev Genet*, 5(2) :101–113.

[Bauman, 1996] Bauman, S. (1996). A linear algorithm for the homogeneous decomposition of graphs. Technical Report TUM-M9615, Technische Universität München. Editeur : H.Wähling Fakultät für Mathematik der Technishen Universität München, D-80290 München, Germany.

[Gagneur et al., 2004] Gagneur, J., Krause, R., Bouwmeester, T., and Casari, G. (2004). Modular decomposition of protein-protein interaction networks. *Genome Biology*, 5 :R57.

[Gallaï, 1967] Gallaï, T. (1967). Transitiv orientierbare graphen. *Acta Math. Acad. Sci. Hungar.*, 18 :25–66.

[Guimera and Amaral, 2005] Guimera, R. and Amaral, L. A. (2005). Functional cartography of complex metabolic networks. *Nature*, 433(7028) :895–900.

[Holme et al., 2003] Holme, P., Huss, M., and Jeong, H. (2003). Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 19 :532–538.

[Ishihara et al., 2005] Ishihara, S., Fujimoto, K., and Shibata, T. (2005). Cross talking of network motifs in gene regulation that generates temporal pulses and spatial stripes. *Genes to Cells*, 10(11) :1025–1038.

[Jamison and Olariu, 1995] Jamison, O. and Olariu, S. (1995). p-component and the homogeneous decomposition of graphs. *SIAM Journal of Discrete Mathematics*, 8 :448–463.

[Kell and Knowles, 2006] Kell, D. B. and Knowles, J. D. (2006). The role of modeling in systems biology : From concepts to nuts and bolts. pages 3–18.

[McConnell and Spinrad, 1994] McConnell, R. M. and Spinrad, J. P. (1994). Linear-time modular decomposition and efficient transitive orientation of comparability graphs. In *SODA '94 : Proceedings of the fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 536–545, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

[Paul, 2006] Paul, C. (2006). Aspect algorithmique de la décomposition modulaire. HDR.

[Ravasz and al, 2002] Ravasz, E. and al (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297 :1551–1555.

[Rives and Galitski, 2003] Rives, A. W. and Galitski, T. (2003). Modular organisation of cellular networks. *PNAS*, 100(3) :1128–1133.

[Schuster et al., 2002] Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I., and Dandekar, T. (2002). Exploring the pathway structure of metabolism : Decomposition into subnetworks and application to mycoplasma pneumoniae. *Bioinformatics*, 18 :351–361.

[Szallasi et al., 2001] Szallasi, Z., Periwal, V., and Stelling, J. (2001). *On modules and modularity : From system modeling in cellular biology*. MIT Press.

[Szerlong et al., 2003] Szerlong, H., Saha, A., and Cairns, B. (2003). The nuclear actin-related proteins arp7 and arp9 : a dimeric module that cooperates with architectural proteins for chromatin remodeling. *EMBO*, 22 :3175–3187.

[Wang and Zhang, 2007] Wang, Z. and Zhang, J. (2007). In search of the biological significance of modular structures in protein networks. *PLoS Computational Biology*, 3(6) :e107 (1–11).

[Zotenko et al., 2006] Zotenko, E., Guimaraes, K. S., Jothi, R., and Przytycka, T. M. (2006). Decomposition of overlapping protein complexes : A graph theoretical method for analyzing static and dynamic protein associations. *Algorithms for Molecular Biology*, 1(1) :7.

Sommaire

Décomposition homogène des réseaux macromoléculaires

Del Mondo Géraldine
(encadrée par Eveillard Damien et Rusu Irena)

Résumé

Il est aujourd'hui accepté que les systèmes biologiques sont composés d'unités fonctionnelles[Rives and Galitski, 2003][Guimera and Amaral, 2005]. Ces unités fonctionnelles sont communément perçues comme des entités semi-autonomes qui présentent des connections fonctionnelles denses avec les autres unités et plus laches avec l'environnement. De telles unités sont impliquées à divers niveaux d'organisation du vivant. On souhaite retrouver ces unités fonctionnelles en utilisant une description modulaire des réseaux d'interaction macromoléculaire. Cette description est devenue aujourd'hui nécessaire pour appréhender la complexité des systèmes biologiques.

Termes généraux : décomposition modulaire, décomposition homogène, module, réseau macromoléculaire, théorie des graphes

LINA, Faculté des Sciences & Techniques de Nantes
2, rue de la Houssinière
B.P. 92208 — F-44322 NANTES CEDEX 3