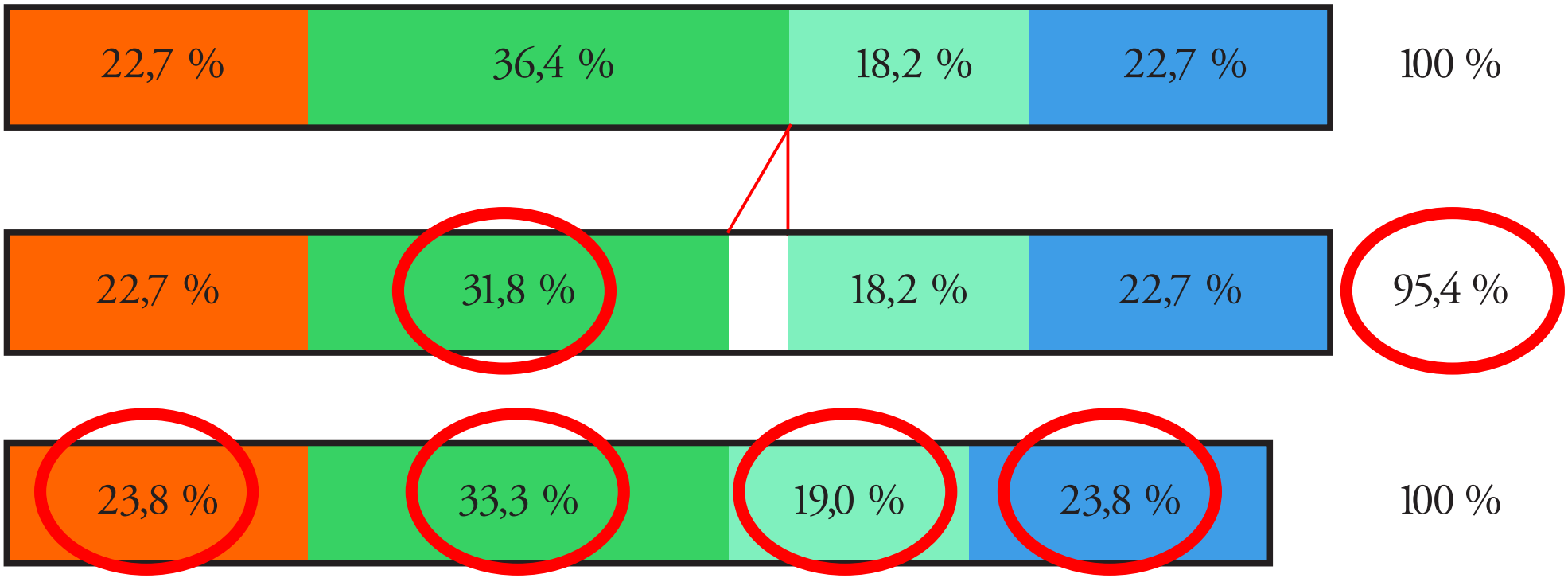
Analyse de données compositionnelles par recherche de sous-graphes disjoints

Emmanuel Curis

J. G. A., novembre 2020

Introduction — Données compositionnelles ①

- Données sur la composition d'un système
  - ⇨ K constituants, 1 à K — q_i : quantité du i-ème
- Exprimées en fraction d'un tout



- ⇨ Elles sont corrélées : somme imposée !
- ⇨ Changer l'une modifie toutes les autres

Introduction — Données compositionnelles ②

Applications en chimie, géologie, archéologie...

- Composition d'un mélange : fractions molaires, massiques
  - ⇨ Composition des minéraux en oxydes (20 % de SiO_2...)
  - ⇨ Composition des huiles en acides gras

Applications en biologie...

- Des situations « évidentes »...
  - ⇨ pourcentage de leukocytes de chaque type
- ... et d'autres plus inattendues !
  - ⇨ Données d'expression (quantification des A. R. N.)
  - ⇨ Histologie des champs microscopiques...

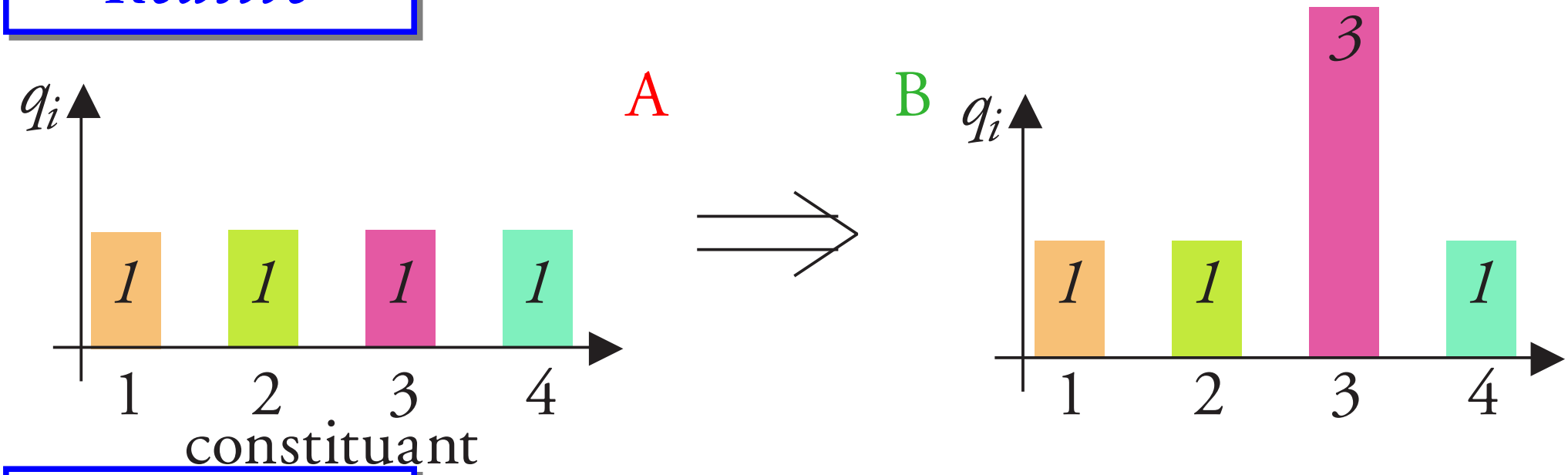
Les notations...

- K : nombre total de constituant
  - ⇨ i, j : indice identifiant le constituant
- q_i : quantité (absolue) du constituant i
  - ⇨ masse (g), quantité de matière (mol)...
- M : quantité (masse) totale
- x_i : proportion du constituant i dans le total
  - ⇨ fraction molaire, fraction massique...
  - ⇨ $x_i = \dfrac{q_i}{\sum_{j=1}^{K} q_j}$
- On utilisera aussi $\ln x_i$

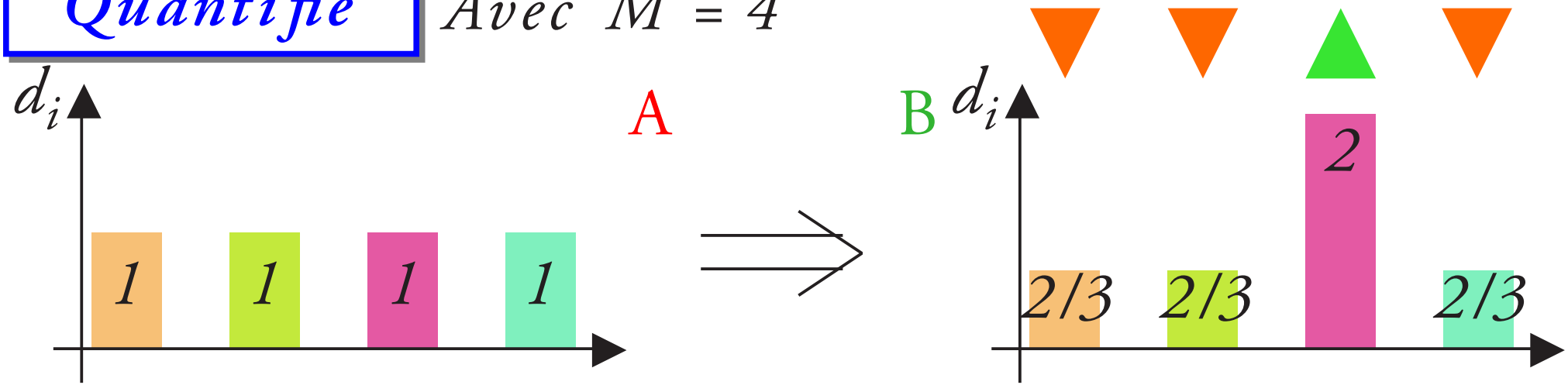
Conséquences sur les conclusions — Exemple

★ K = 4 constituants différents, 2 conditions A & B

Réalité



Quantifié Avec $M = 4$



Rationnel : les rapports ne sont pas faussés

- Soient deux composants, i (q_i, x_i) et j (q_j, x_j)
- Rapport des quantités réelles : $r_{i,j} = \dfrac{q_i}{q_j}$
- Rapport des quantités relatives :

$$r^*_{i,j} = \frac{x_i}{x_j} = \frac{q_i}{\sum_{k=1}^{K} q_k} \Bigg/ \frac{q_j}{\sum_{k=1}^{K} q_k} = \frac{q_i}{q_j} = r_{i,j}$$

- L'information sur les quantités relatives persiste
  - ⇨ Peut être celle cherchée (équilibre chimique…)
- Comment interpréter cette information sinon ?
  - ⇨ Que signifie un changement du rapport moyen ?

Interpréter les modifications des rapports ①

- Exemple : 2 composants, 3 changements possibles chacun

composant j \ composant i	*Inchangé*	*Doublé*	*Divisé par 2*
Inchangé	$r_{i,j}$ inchangé	$r_{i,j}$ doublé	$r_{i,j}$ divisé par 2
Doublé	$r_{i,j}$ divisé par 2	$r_{i,j}$ inchangé	$r_{i,j} \times 1/4$
$\times$ *1/2*	$r_{i,j}$ doublé	$r_{i,j} \times 4$	$r_{i,j}$ inchangé

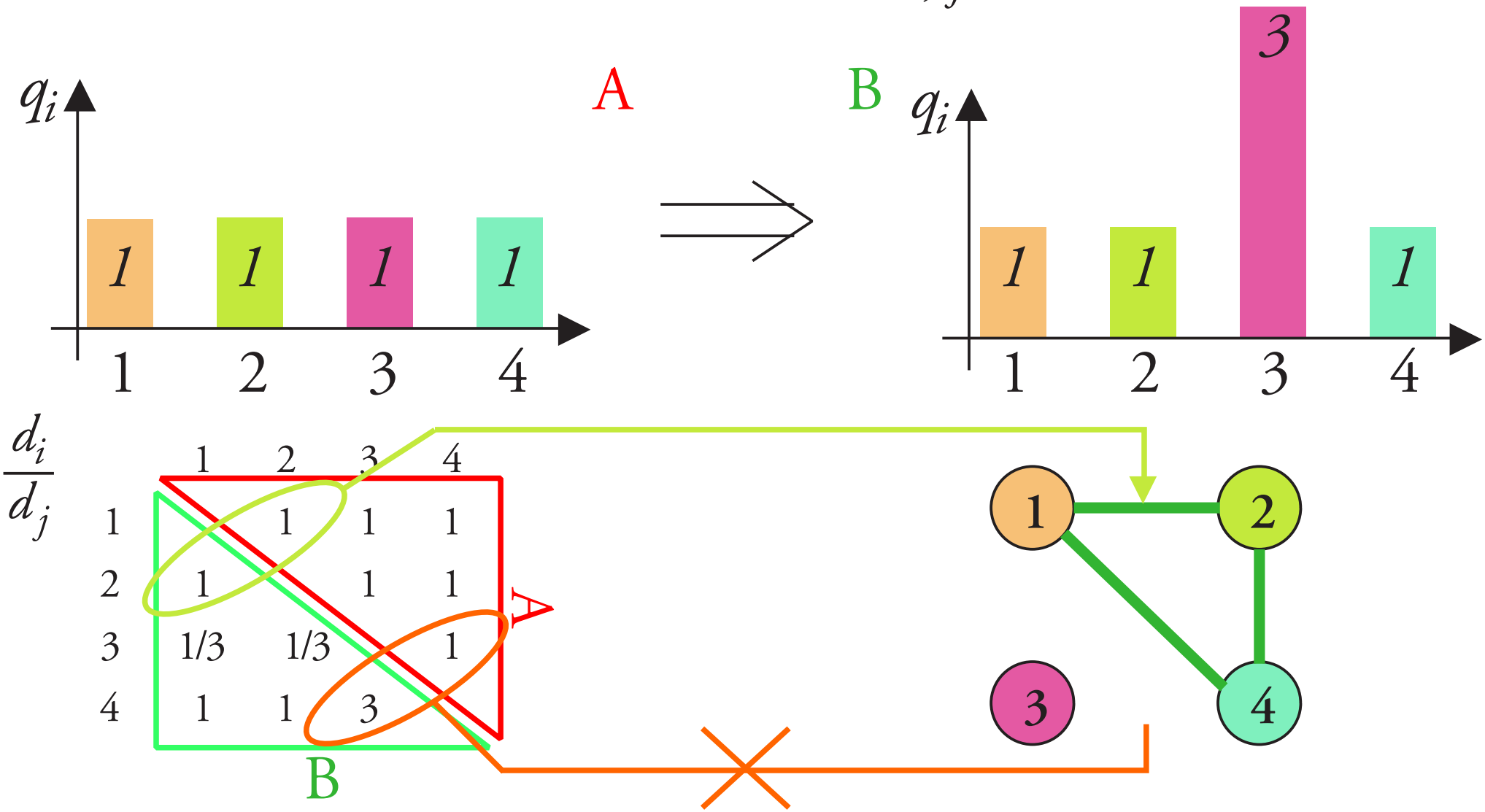
Le rapport est inchangé

- Les quantités des deux composants sont inchangées…
- … ou les deux ont été modifiées par le même facteur.

Si le passage de A à B ne modifie pas $r_{i,j}$, alors il a le même effet sur les quantités des composants i et j.

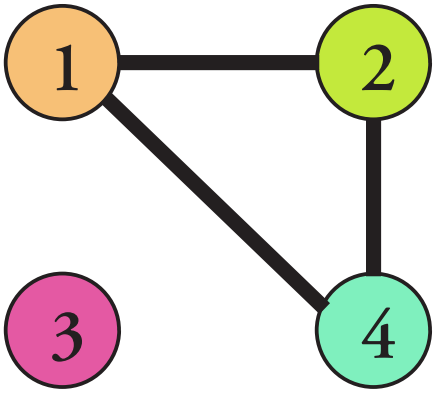
Construire un graphe des composants quantifiés

- Nœuds du graphe : les K* composants quantifiés
- Les nœuds i et j sont reliés ssi $r_{i,j}$ est inchangé



Interpréter les changements des rapports ②

- Plusieurs sous-graphes disjoints
  - ⇨ Chacun est complètement connexe



- Chaque sous-graphe correspond à une variation différente
  - ⇨ Au plus un pour « pas de changement »
  - ⇨ Mais impossible de savoir lequel…

Limitation des données compositionelles

- En pratique, les résultats ne seront pas aussi tranchés
  - ⇨ Connexions indues entre les nœuds…
  - ⇨ Connexions absentes entre les nœuds…

Comment construire le graphe empirique ? ①

- Les nœuds sont connus : composants quantifiés

- En revanche, les arêtes ne le sont pas
  - ⇨ Il faut une règle pour mettre ou pas les arêtes
  - ⇨ Deux approches possibles : constructive et destructive
  - ⇨ Dans les deux cas, il faut tester toutes les arêtes…

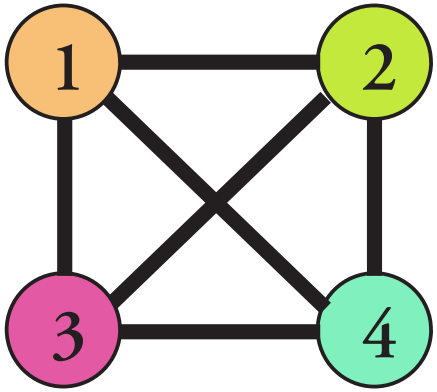
Comment construire le graphe empirique ? ②

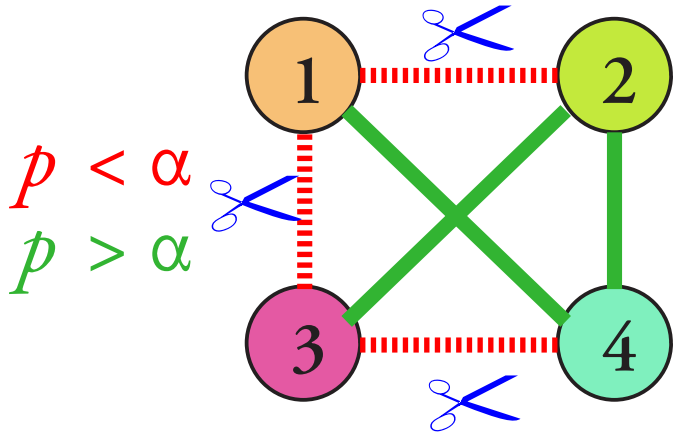
Approche « destructive »

- On part d'un graphe complet
- On supprime l'arête entre deux constituants si l'on prouve qu'ils se comportent différemment

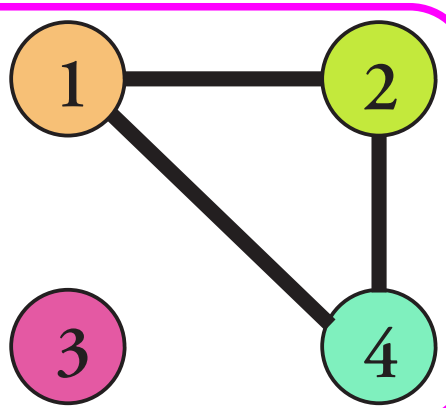
⇨ Utilisation de tests de différence, « classiques »

Départ





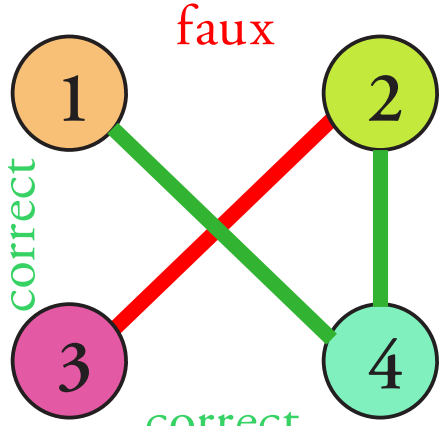
Réalité



Résultat

Incorrect

Correct

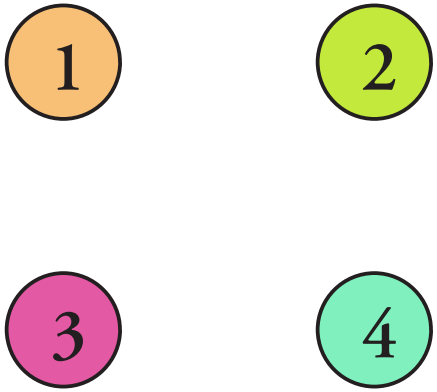


Comment construire le graphe empirique ? ③

Approche constructive

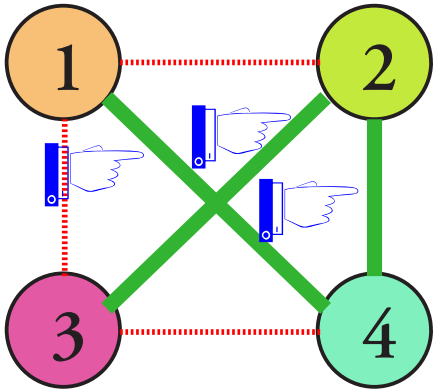
- On part d'un graphe sans aucune arête

Départ



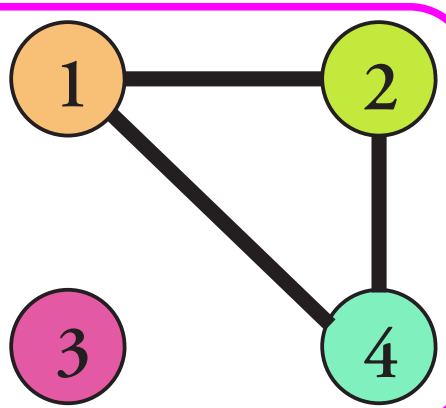
- On ajoute une arête entre deux constituants si l'on prouve qu'ils se comportent (suffisamment) pareil

$p > \alpha$
$p < \alpha$

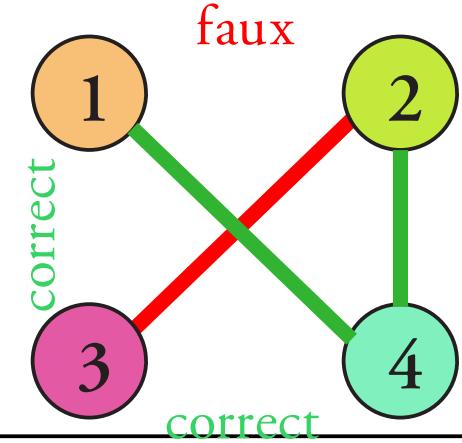


⇨ Utilisation de tests d'équivalence

Réalité



Résultat
Incorrect
Correct



Déterminer si un rapport a été modifié

- ***Étape clef de la méthode***
  - ⇨ Détermine la structure du graphe !
- Estimer la variation du rapport
  - ⇨ Modèle statistique adapté au plan expérimental
  - ⇨ Variation cherchée : l'un des paramètres du modèle, θ
  - ⇨ Typiquement : modèle log-linéaire
- Les nœuds i et j sont déliés si $r^{\mathrm{B}}_{i,j} / r^{\mathrm{A}}_{i,j}$ est significativement différent de 1
  - ⇨ Dans le modèle, si le test de θ est significatif
- Quel niveau (« seuil de p ») utiliser pour ce test ?

Déterminer des groupes de composants

Comment gérer les fausses connexions ?

- Ne considérer que les sous-graphes disjoints
  - ⇨ même s'il manque des connexions dedans
  - ⇨ très sensible aux variations non-détectées
- Ne considérer que les ensembles connexes de nœuds
  - ⇨ cliques et cliques maximales
  - ⇨ très sensible aux fausses variations
  - ⇨ calculs longs pour les grands graphes (RNAseq…)
- Recherche de communautés
  - ⇨ Plusieurs définitions & algorithmes…

Choix du niveau du test — sous-graphes disjoints ①

- Observation : « Au moins deux graphes disjoints »
  - ⇨ Si aucun changement, doit arriver avec prob. $< \alpha$ (H_0)
  - ⇨ Quel niveau α_0 utiliser dans le test du rapport ?

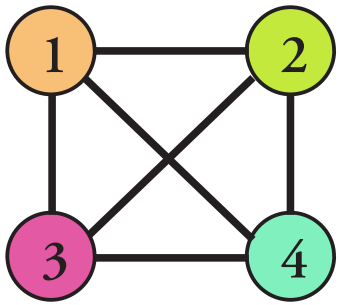
- Se produit si un nœud (le 1) n'a aucune connexion
  - ⇨ Si *tous* les rapports $r_{1,\,j}$ sont significativements modifiés
  - ⇨ Aucune correction de multiplicité nécessaire

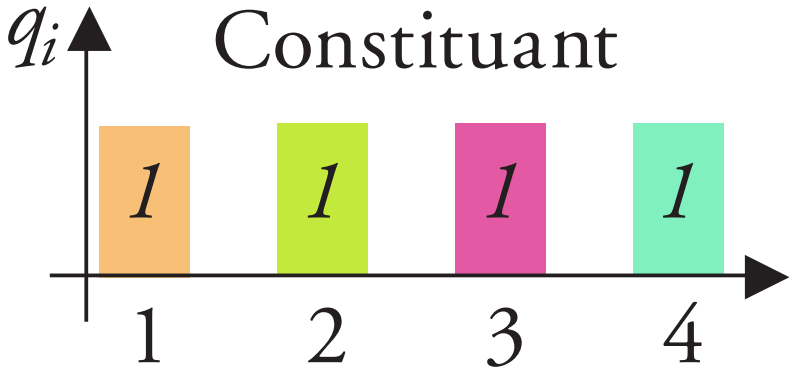
- Si les tests étaient indépendants, $\Pr(\text{TSM}|H_0) = \alpha = \alpha_0^{K^*-1}$
  - ⇨ α_0 doit être (bien) plus grand que α

- Les tests ne sont **pas** indépendants. Et doit être vrai pour tous les nœuds…

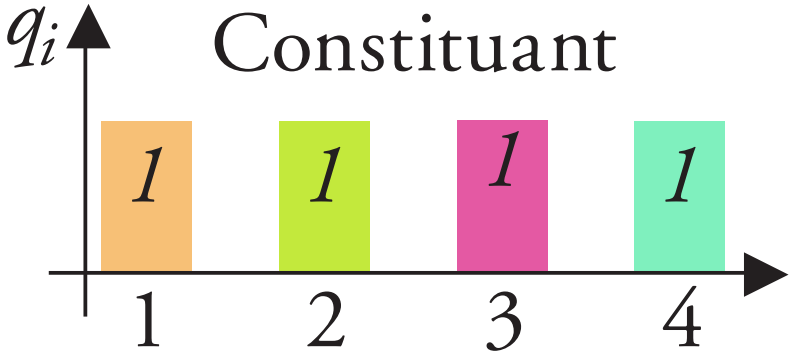
Étude sous H_0 — Principe ①

Réalité





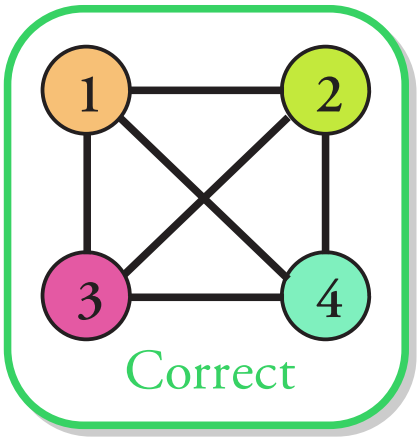
A ⟹ B

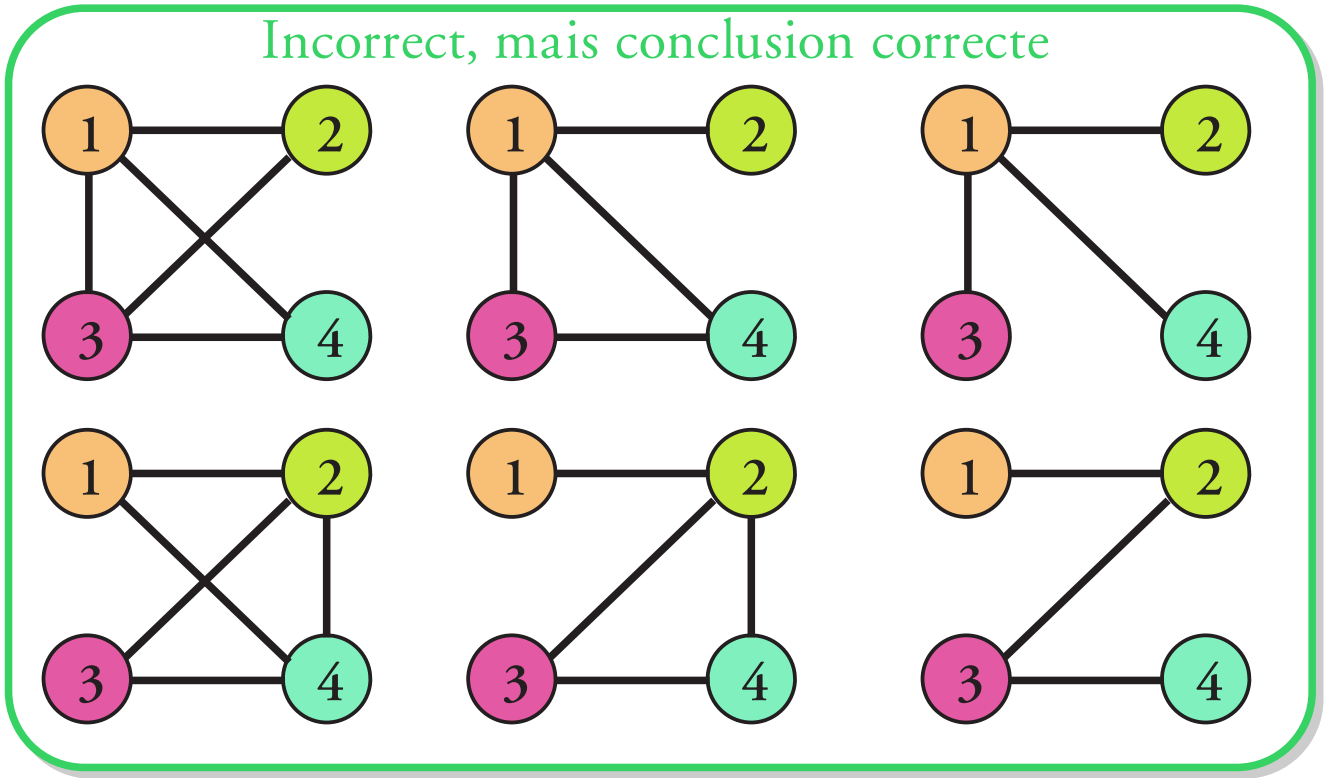


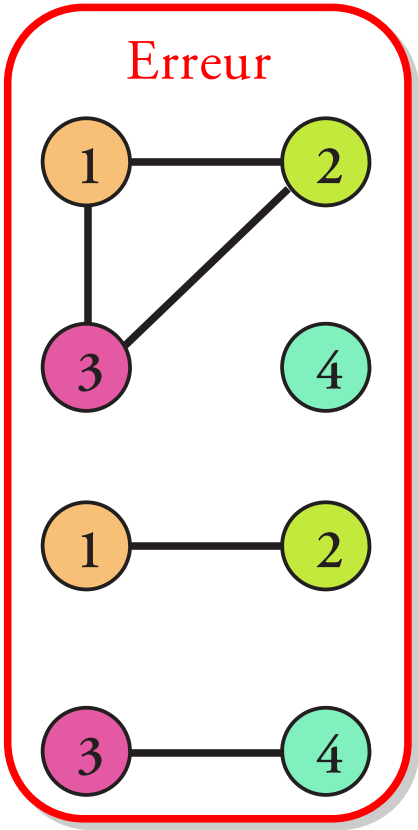
- Distribution log-normale des q_i ; CV = 20 %
  - ⇨ Test de Student « classique », en ln, pour chaque arête
- Sous H_0, la taille des groupes joue peu

Étude sous H_0 — Principe ②

- Sous H_0, rien ne se passe entre A et B
  - ⇨ On ne doit obtenir qu'un seul graphe
  - ⇨ Le pourcentage de simulations se trompant estime α



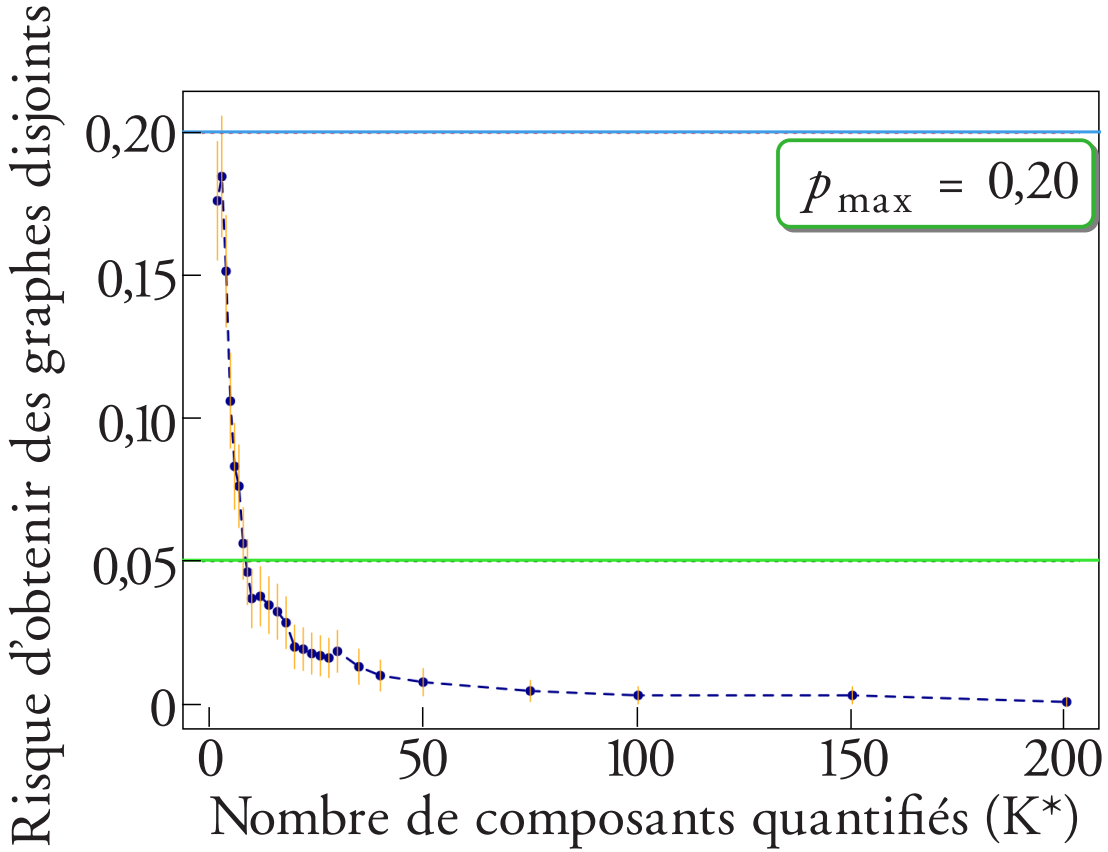


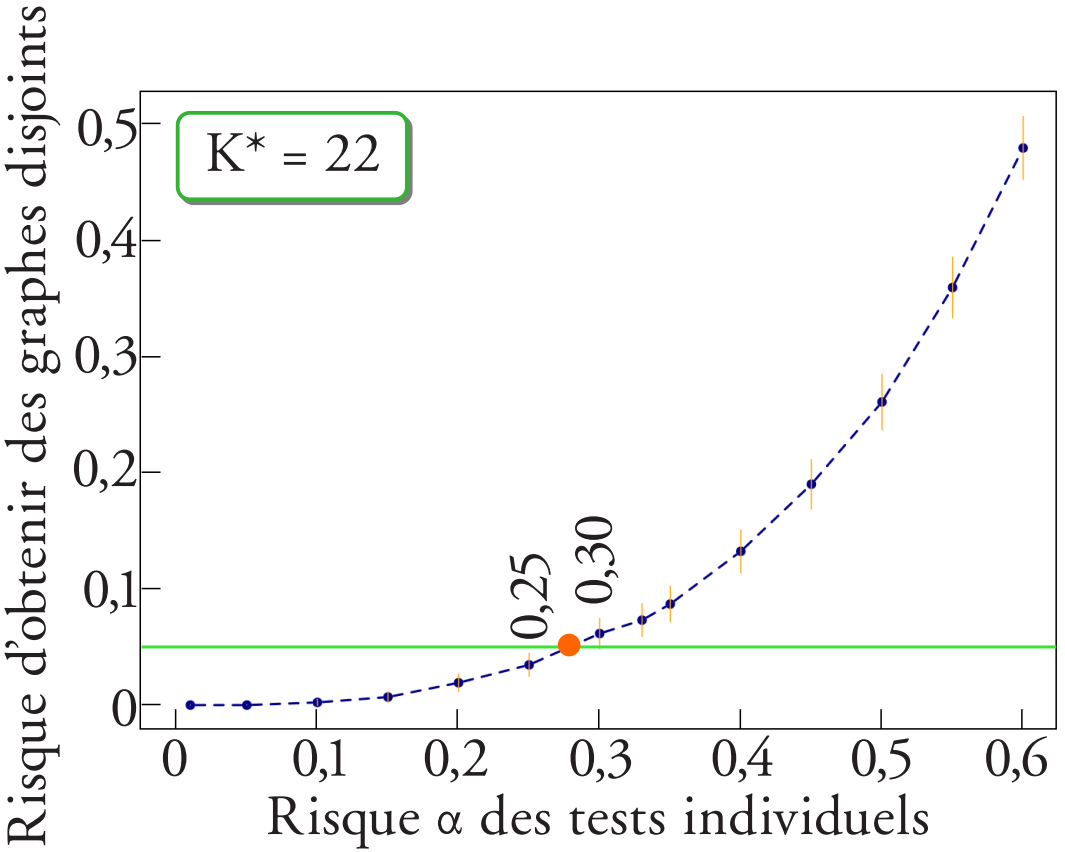


Choix du niveau du test — graphes disjoints ②

Résultats de simulation

- Valeurs log-normales, sous H_0, somme valant 1
- 10 000 simulations, avec K = 200 composants, 2 groupes





Et pour la puissance ?

- Simulations précédentes : risque de détecter des composants se comportant différemment, quand tous se comportent de façon identique (« rejet de H_0 quand elle est vraie »)

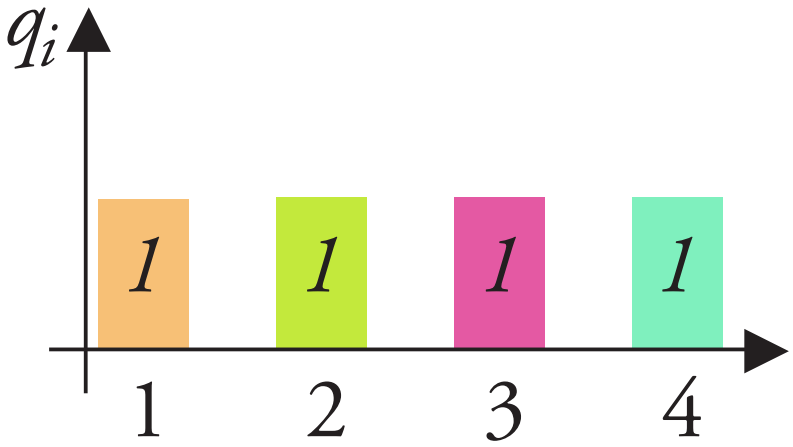
- Comment se comporte la méthode quand certains composants se comportent réellement différemment ?
  - ⇨ Détecte-t-elle des graphes disjoints ? (*puissance*)
  - ⇨ Détecte-t-elle les bons groupes de composants ?

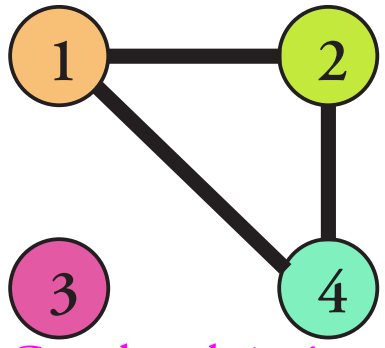
- Dépend de la puissance de chaque test individuel

Étude de l'exemple — conditions de simulation

- K = 4 composant, 2 conditions A & B, 1 composant triple

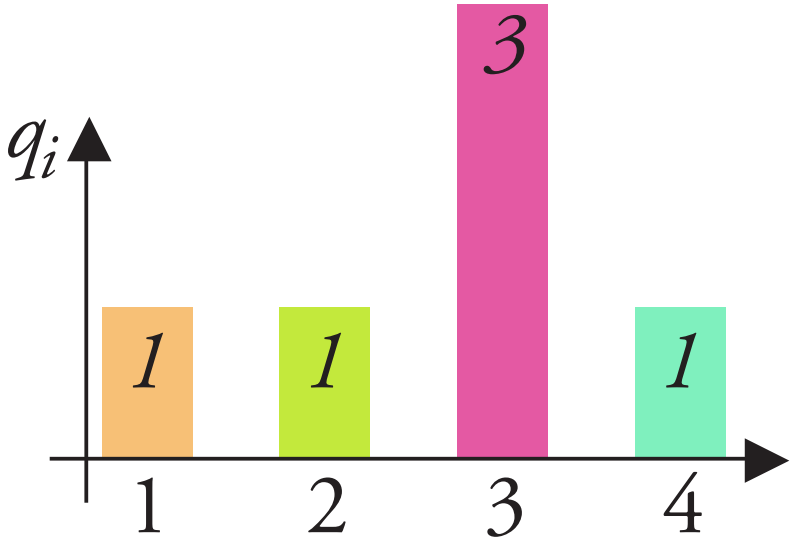
Réalité





Graphe théorique

A ⇒ B

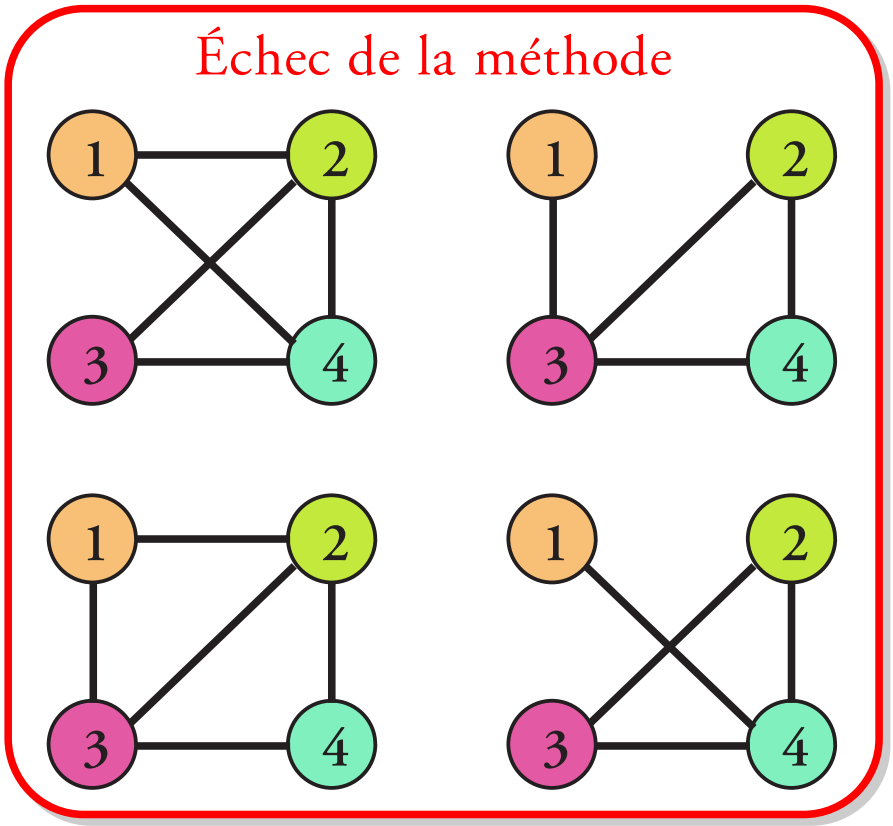


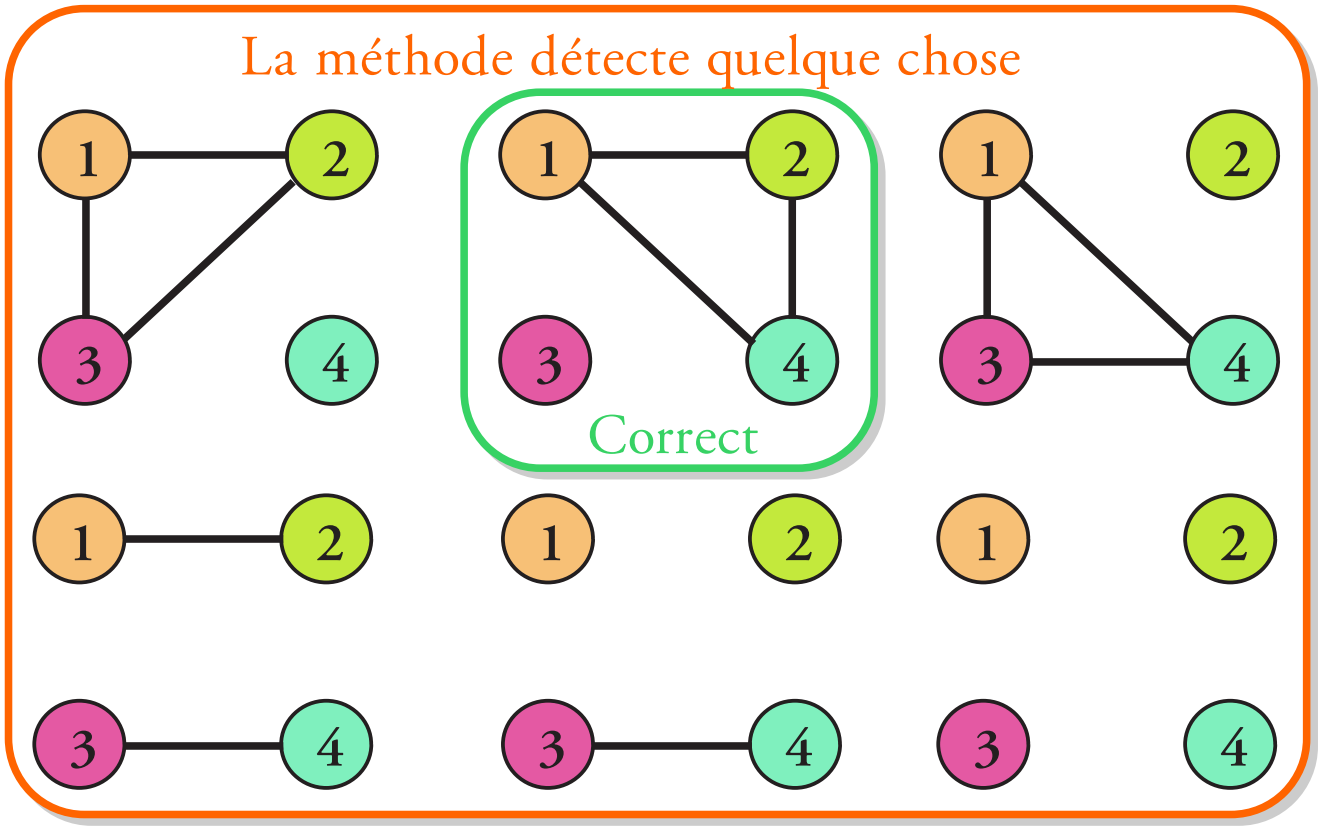
- Distribution log-normale ; CV = 20 %

Étude de l'exemple 1 — étude de la puissance totale

- On rejette H_0 pour n'importe quelle raison
  - ⇨ Il se passe quelque chose (ce qui est observé… ou pas !)
  - ⇨ Graphe disjoint, quel qu'il soit

- Mais est-ce que l'on détecte bien ce qu'il faut ?





Exemples d'application (classés par K croissant)

Le Cas général (K > 2) — K « très faible »

K = 6 — Composition tissulaire ①

Données d'Anne-Gaëlle Cordier

- Étude microscopique de placentas humains
  - ⇨ Six types de structures repérés sur les champs
  - ⇨ Ces 6 types recouvrent tout le champ
  - ⇨ Pour chaque placenta, 5 champs sont étudiés
- Influence de la drépanocytose sur l'importance de ces structures
  - ⇨ groupe de placentas de femmes témoin
  - ⇨ groupe de placentas de femmes drépanocytaires
  - ⇨ groupes équilibrés, $n = 7$ femmes par groupe

K = 6 — Composition tissulaire ②

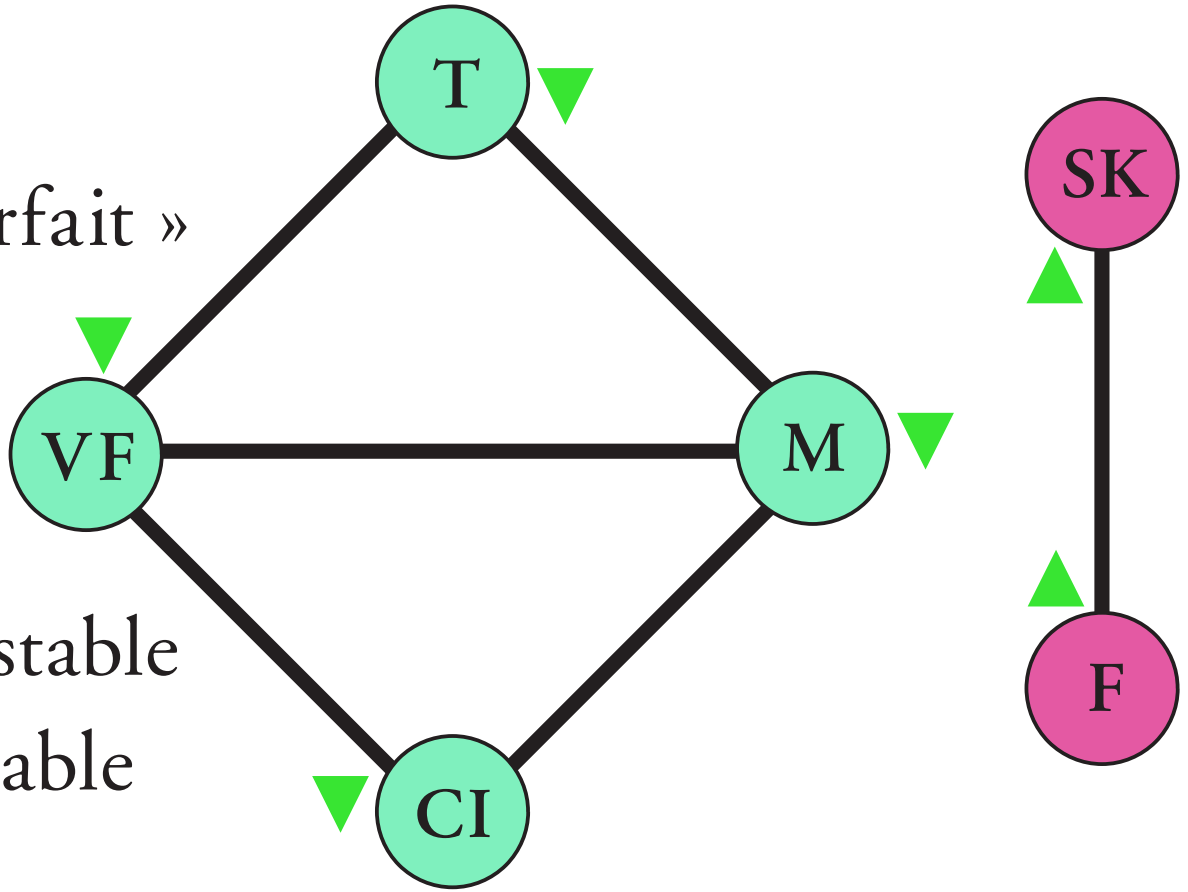
- Quantification : surface occupée par le tissu dans le champ

- Le champ est (beaucoup) plus petit que le placenta
  - ⇨ la somme des surfaces mesurées vaut forcément celle du champ
  - ⇨ données compositionnelles par contrainte expérimentale !

- K = 6 « composants » — T, VF, M, SK, F et CI
  - ⇨ D'après les simulations précédentes, avec $n = 7$, le seuil de coupure est à $p < 0{,}167$
  - ⇨ Intervalle de confiance de ce seuil : [0,159 ; 0,175]
  - ⇨ $6 \times 5 / 2 = 15$ arêtes à tester

K = 6 — Composition tissulaire ③

- Coupure : $p < 0{,}16$
- Graphe obtenu « quasi-parfait »
  - ⇨ Une arête manque…
- Deux groupes nets
  - ⇨ Un augmente ou reste stable
  - ⇨ Un diminue ou reste stable



Le Cas général ($K > 2$) — K « faible »
Application à la RT-qPCR

Pourquoi des données compositionnelles ?

- Des cellules sont isolées, mise en culture…
  - ⇨ K A. R. N. différents, [ARN i] = q_i
- Les A. R. N. sont extraits de la culture
  - ⇨ K A. R. N. différents, [ARN i] = q_i
- On isole une masse totale M d'A. R. N. pour quantification
  - ⇨ K A. R. N. différents, [ARN i] = $x_i = \mathrm{M} \dfrac{q_i}{\sum_{j=1}^{\mathrm{K}} q_j}$
- K* < K A. R. N. différents sont quantifiés
  - ⇨ K* A. R. N. différents, [ARN i] = $d_i = \lambda_i x_i = \lambda_i \mathrm{M} \dfrac{q_i}{\sum_{j=1}^{\mathrm{K}} q_j}$

K = 60 — gènes de référence en RT-qPCR ①

- Recherche de gènes « de référence » pour analyse des résultats de RT-qPCR
  - ⇨ Gènes dont la quantité est invariante d'une condition à l'autre

- 2 groupes de patients : témoins (n = 19) et bipolaires (n = 19)
  - ⇨ Leurs lymphocytes sont cultivés

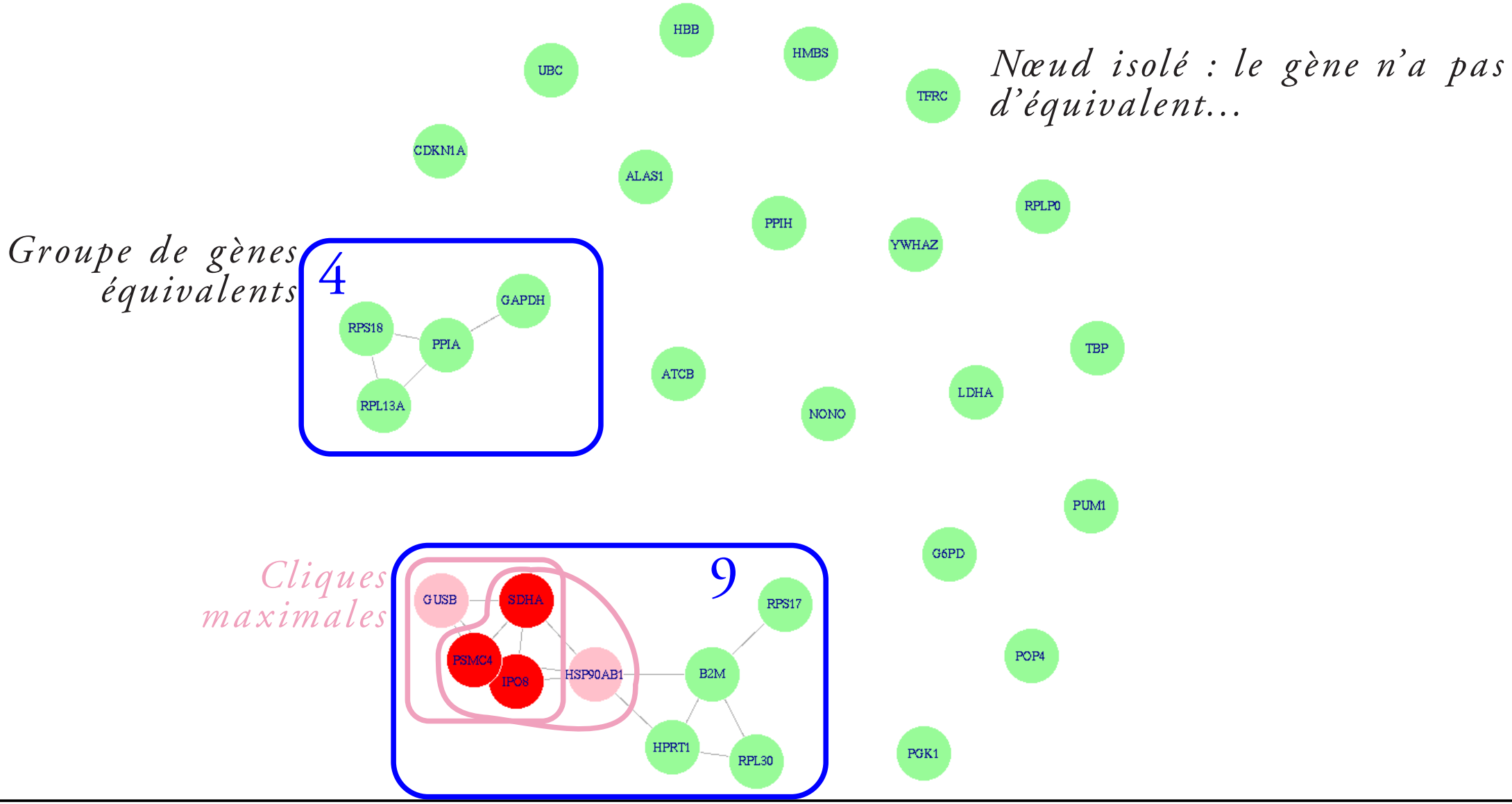
- 60 gènes candidats

- Zone d'équivalence : $\Delta = 0{,}5\ C_t$
  - ⇨ Variabilité technique : de l'ordre de 0,3 (échelle C_t)
  - ⇨ Si efficacité parfaite : ×1,414 au maximum

Curis et coll., novembre 2019

K = 60 — gènes de référence en RT-qPCR ②

- Critère : cliques maximales
- K* = 60 nœuds ⇨ p_{max} = 0,25 pour avoir $\alpha < 0{,}05$



Le Cas général (K > 2) — K « moyen »
Application à la métabolomique

Pourquoi des données compositionnelles ?

- Des échantillons sont prélevés, traités…
  - ⇨ K composés (C) différents, $[C_i] = q_i$
- Les composés sont séparés puis dosés
  - ⇨ K' composés sont séparés, $[C_i] = \varepsilon_i\, q_i$ (avec $\varepsilon_i \leqslant 1$)
  - ⇨ K* composés sont quantifiés, $d_i = \lambda_i\, \varepsilon_i\, q_i$
  - ⇨ Courbes d'étalonnage inconnues : d_i en unité arbitraire

- On normalise les valeurs entre expériences
  - ⇨ La normalisation rend les données compositionnelles

K = 491 — Maladie du greffon contre l'hôte ①

Données de G. SOCIÉ (hôpital Saint-Louis)

- 2 groupes de patients, ayant reçu une greffe de moëlle : patients développant ou ne développant pas cette complication
- Chez chaque patient, prélèvement plasmatique
- Quantification « lipidomique » des petites molécules

Modèle

$$\ln r_{i,j} = \underbrace{\mu_{0,i,j}}_{\text{Rapport moyen en l'absence de complication}} + \underbrace{\delta_{\mathrm{D}\,i,j}}_{\text{Variation en présence de complication}} \mathbf{1}_{\mathrm{D}} + \varepsilon$$

K = 491 — Maladie du greffon contre l'hôte ③

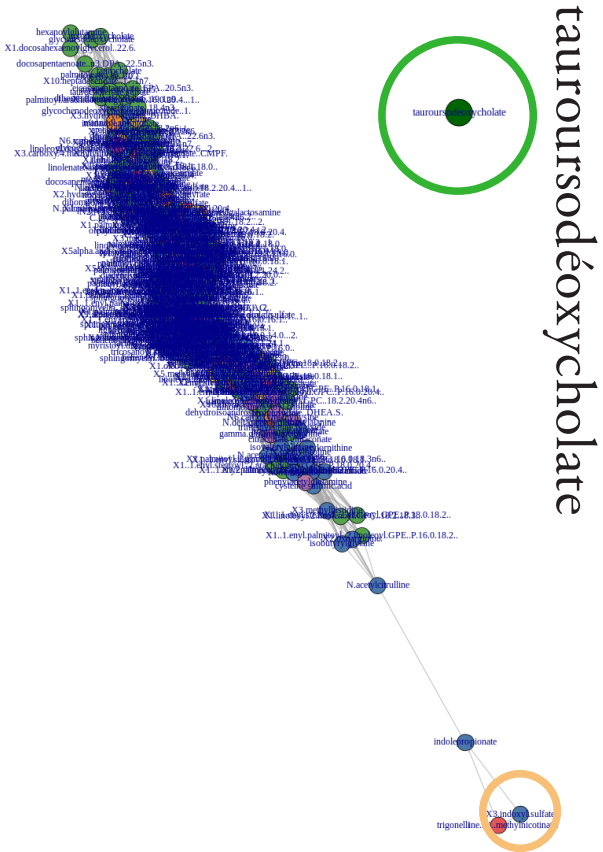
★ Analyse sur 491 métabolites (communs aux deux cohortes)

⇨ Seuil de coupure : $p < 0{,}44$

Michonneau et coll., Nature Communications, décembre 2019

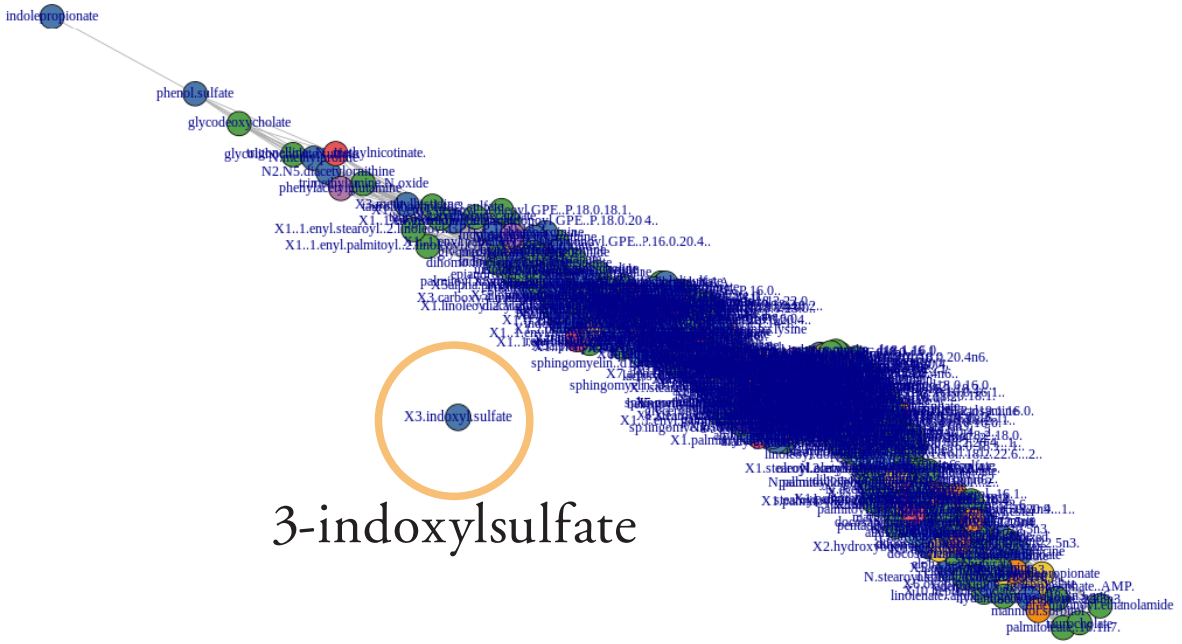
Cohorte 1

12 *vs* 31 patients



Cohorte 2

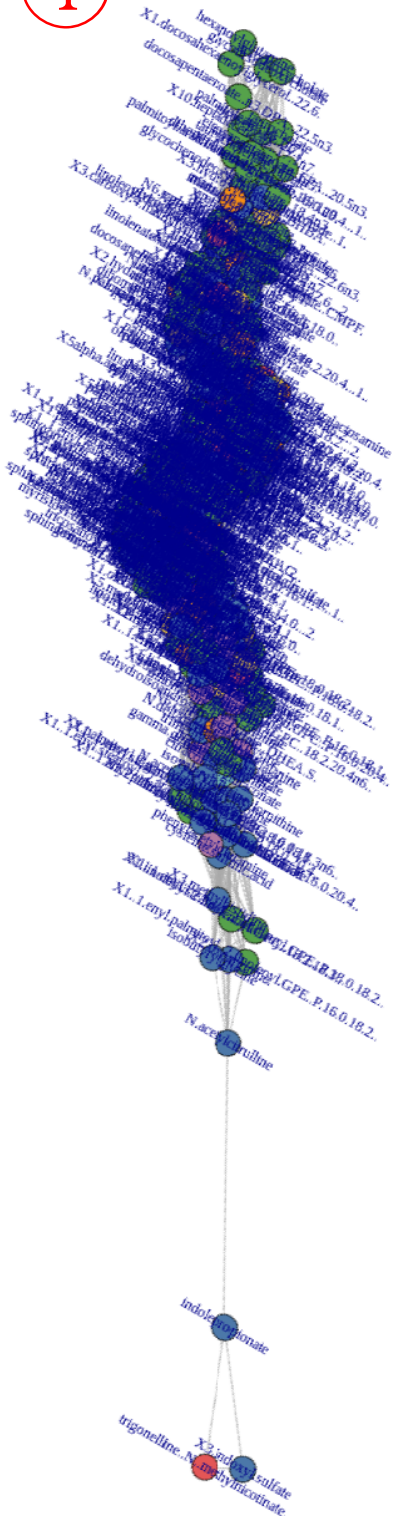
26 *vs* 30 patients



3-indoxylsulfate

K = 491 — Maladie du greffon contre l'hôte ④

- Tous les composés bougent un peu
  - ⇨ Quand K* augmente, le graphe tend à devenir « filaire »
  - ⇨ Certains composés ont une variation intermédiaire entre celles de deux autres
  - ⇨ Difficile de trouver des graphes disjoints
- Construction alternative : distances extrêmes
  - ⇨ Les composés « diamétralement opposés » se comportent différemment
  - ⇨ Étude de la distribution de la distance minimale entre deux nœuds



Quelques limitations de la méthode...

- Nombre de tests augmente avec K* comme K^{*2}
  - ⇨ Temps et mémoire nécessaires augmentent « vite »
  - ⇨ Temps de calcul pour K 800 : 15 minutes
  - ⇨ Temps de 1000 simulations, 48 cœurs : 5 heures
  - ⇨ Peut être problématique en RNA-Seq ($K^* \approx 2 \times 10^4$)
- Taille du graphe augmente avec K*
  - ⇨ Temps d'analyse du graphe peuvent devenir longs (cliques, communautés...)
- Impossible de savoir comment varie un composé donné
  - ⇨ Limite de la normalisation, pas de la méthode...

… et quelques avantages

- Pas besoin d'hypothèse sur des composés invariants
- Pas de correction de multiplicité
  - ⇨ Moins de perte de puissance quand K* augmente…
- Insensible aux différences d'efficacité de quantification entre composés
  - ⇨ tant que ces différences ne dépendent pas des conditions comparées !
- RNA-Seq : insensible aux profondeurs de séquençage…
- Applicable à tout plan expérimental

Bibliographie

Logiciel

- *Package* R : SARP.compo, disponible sur le C. R. A. N.

Articles

Théorie

- E. Curis et coll., *Bioinformatics*, janvier 2019 (approche destructive)
- E. Curis et coll., *Scientific Reports*, novembre 2019

Application

- P. A. Geoffroy et coll., *The World Journal of Biological Psychiatry*, 2018

Merci de votre attention !

Remerciements

- Bruno Blanchet — idée d'un graphe…
- Charles-Henry Cottart — données hépatiques (traceur)
- Anne-Gaelle Cordier — données d'imagerie (placentas)
- Cindie Courtin, Calypso Nepost, Pierre-Alexis Geoffroy — RT-qPCR
- David Michonneau, Gérard Socié — données métaboliques
- Bruno Saubaméa — données de RNA-Seq
- Étudiants de l'ENSAI
- Sylvie Chevret, Yves Rozenholc, Chantal Guihenneuc