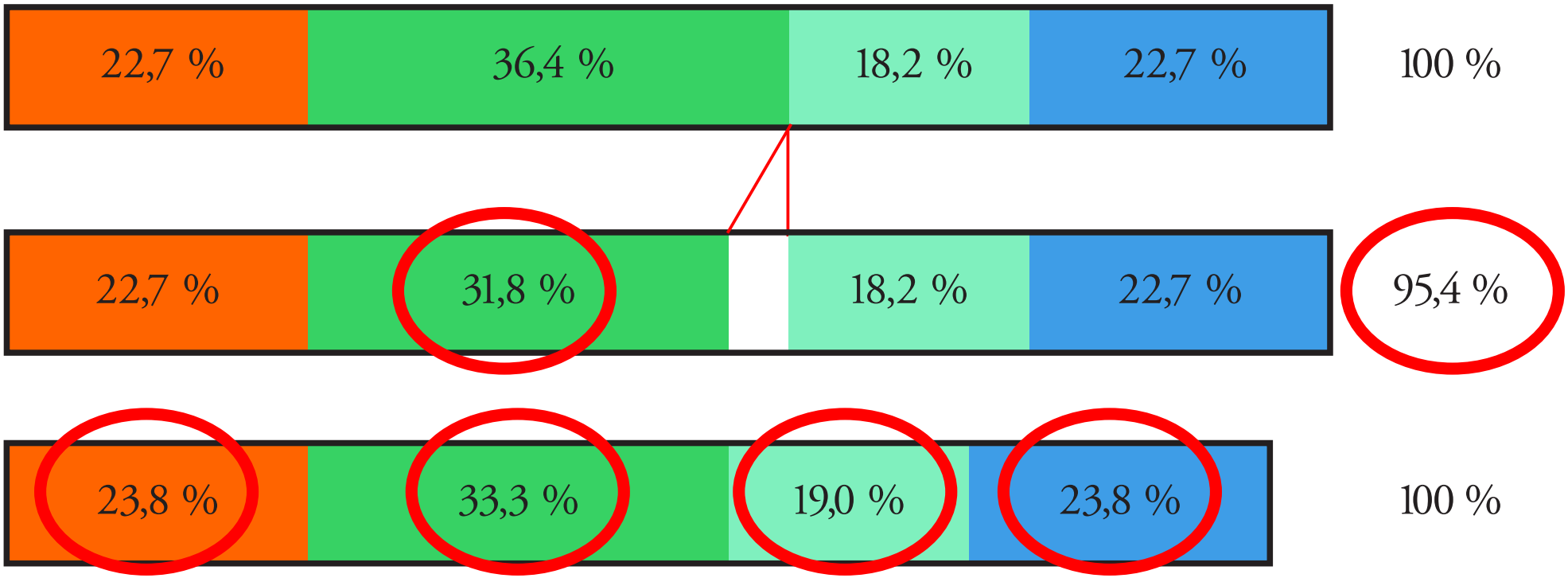
Analyse de données compositionnelles par recherche de sous-graphes disjoints

Emmanuel Curis, Cynthia Marie-Claire

Séminaire du C. N. A. M. — vendredi 1er mars 2019

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...)

Applications en biologie...

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

Introduction — Origine de ces données

Contrainte intrinsèque du système...

- Une quantité prédéfinie de matière se répartit entre diverses possibilités
  - ⇨ Spéciation d'un élément en quantité totale imposée
  - ⇨ Traceur entre divers organes, tissus...

Contrainte expérimentale « externe »

- Chaque constituant est libre de varier
- Le dosage utilise une quantité imposée du mélange
  - ⇨ Dosage de fractions, plus de quantités
  - ⇨ A. R. N., protéines...
  - ⇨ Populations lymphocytaires

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)...
- x_i : proportion du constituant i dans le total
  - fraction molaire, fraction massique...
  - $$x_i = \frac{q_i}{\sum_{j=1}^{K} q_j}$$
- On utilisera aussi (souvent) $\ln x_i$

Le Cas particulier K = 2

Cas particulier : K = 2

- En théorie, on peut analyser au choix l'une des deux composantes (x_1 ou x_2)
  - ⇨ l'autre varie nécessairement dans l'autre sens
  - ⇨ laquelle choisir ?
  - ⇨ puisque contraint entre 0 et 100 %, problèmes sur les bords

- Transformation classique :

$$z = \ln \frac{x_1}{x_2} = \ln \frac{x_1}{1 - x_1} = -\ln \frac{x_2}{1 - x_2}$$

  - ⇨ Une seule analyse à réaliser incluant toutes les données
  - ⇨ Transformation (régression) logistique…

Cas particulier : K = 2

Exemple applicatif : test d'anhédonie

- Dans leur cage, les souris ont le choix entre deux boissons
  - ⇨ une bouteille d'eau plate ;
  - ⇨ une bouteille d'eau sucrée (saccharose).

- On mesure le volume consommé de chaque boisson par la souris pendant une période donnée
  - ⇨ Q_1 : volume d'eau plate ; Q_2 : volume d'eau sucrée.
  - ⇨ Normalement, la souris consomme davantage d'eau sucrée ($Q_2 > Q_1$)
- Souvent exprimé en proportion : $P = Q_2 / (Q_1 + Q_2)$
  - ⇨ Un traitement a-t-il un effet sur cette proportion ?

Cas particulier : K = 2 — Exemple ①

- Représentation sous forme de graphe
  - ⇨ 2 composants : 2 nœuds (sommets), au plus une arête
  - ⇨ Les nœuds sont reliés s'ils se comportent « pareil »

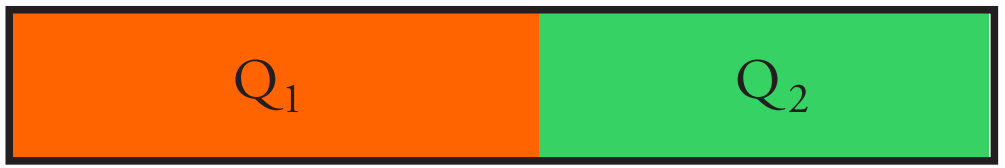
Avant traitement

Volume total de boisson
$Q_1 + Q_2$ (libre)

Après traitement…

… sans effet

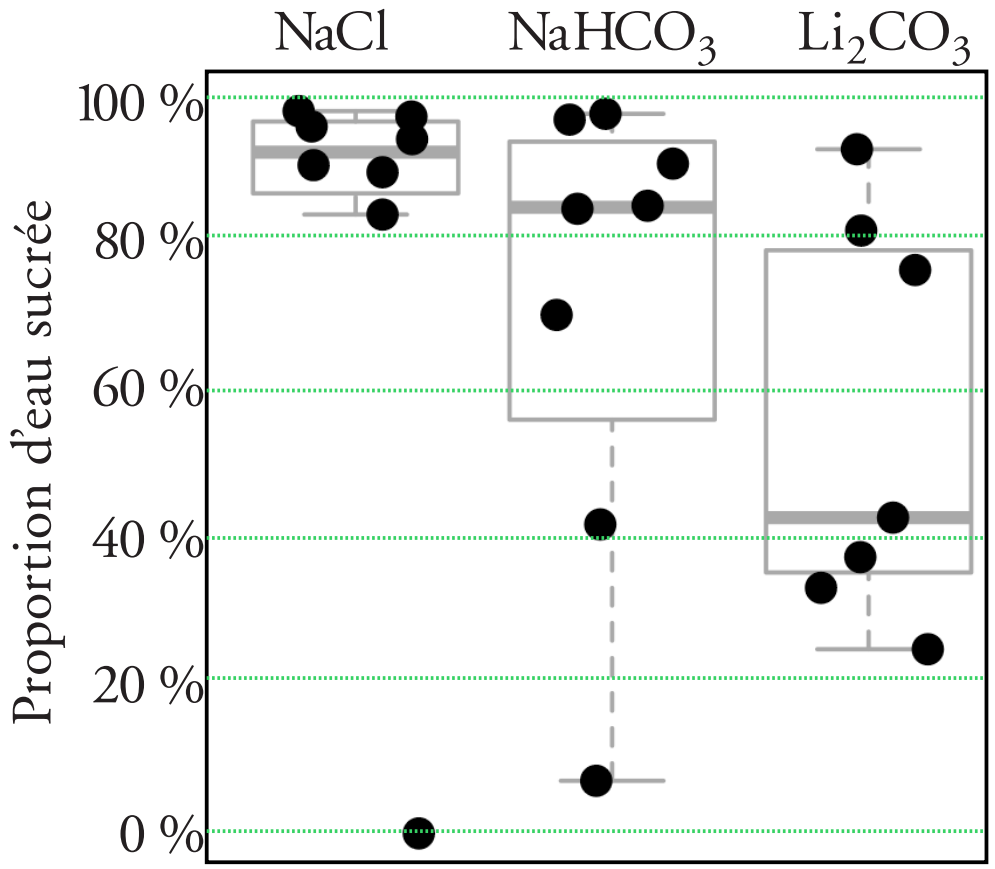
… ayant un effet



Cas particulier : K = 2 — Exemple ②

Expériences d'Anne-Sophie Hannak, UMR-S 1144

- Effet du lithium sur l'état hédonique des rats
  - ⇨ Trois groupes : NaCl, $NaHCO_3$ et Li_2CO_3
  - ⇨ Deux temps : 24 et 48 h — pour les mêmes rats
  - ⇨ 8 rats par groupe
- Données à 24 h utilisées ici
- Résultats évidents
  - ⇨ Un rat atypique (NaCl)
  - ⇨ Effet anhédonique du carbonate
- Et le lithium ?



Cas particulier : K = 2 — Exemple ③

Test global

- Analyse de variance à un facteur
  - ⇨ Sur les proportions : $p = 0{,}0366$
  - ⇨ Sur les $\ln(X_2 / X_1) = \ln(Q_2 / Q_1)$: $p = 0{,}0273$
  - ⇨ Q_1 et Q_2 se comportent différemment

Test pour le lithium

- Test de Student
  - ⇨ Sur les proportions : $p = 0{,}3131$
  - ⇨ Sur les $\ln(X_2 / X_1)$: $p = 0{,}3027$
  - ⇨ Échec à montrer que Q_1 et Q_2 se comportent différemment

Le Cas général (K > 2)

Méthodes classiques (K > 2) — « alr »

- Données brutes non-analysables
  - ⇨ Corrélées, sens de variation fortuit…

Aitchison, 1982 ; 1984…
Egozcue, 2003
Filzmoser, 2009
…

- On choisit un constituant « de référence » (p. ex. le 1)
- On calcule $r_{i,1} = q_i/q_1$ pour les K – 1 autres constituants
- On travaille sur $\ln r_{i,1}$
  - ⇨ Méthode du log des rapport : « alr »
- Les résultats dépendent de la référence choisie
  - ⇨ Comment choisir la référence ?
- Le nombre de tests augmente avec K : perte de puissance !

Méthodes classiques (K > 2) — « clr »

- On pose m_g la moyenne géométrique des K q_i
- On calcule $r^c_{i,c} = q_i / m_g$ pour les K constituants
- On travaille sur $\ln r^c_{i,c}$

  ⇨ Méthode du log des rapport centrés : « clr »
- Les $r^c_{i,c}$ sont toujours compositionnels !
- $r^c_{i,c}$ dépend de tous les constituants : s'il change, pourquoi ? est-ce dû au constituant i ?
- Donne une information globale : « la composition a changé »

Méthodes classiques (K > 2) — « ilr »

- On pose $m_{g,k}$ la moyenne géométrique des k derniers q_i
- On calcule $r_i^i = q_i / m_{g,i+1}$ pour les K – 1 premiers constituants
- On travaille sur $[(K - i)/(K - i + 1)]^{0,5} \ln r_i^i$

  ⇨ Méthode du log des rapport isométrique : « ilr »
- Transformation complexe qui assure l'indépendance
- r_i^i dépend des K – i constituants suivants : s'il change, pourquoi ? est-ce dû au constituant i ?

  ⇨ Information globale : « la composition a changé »

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

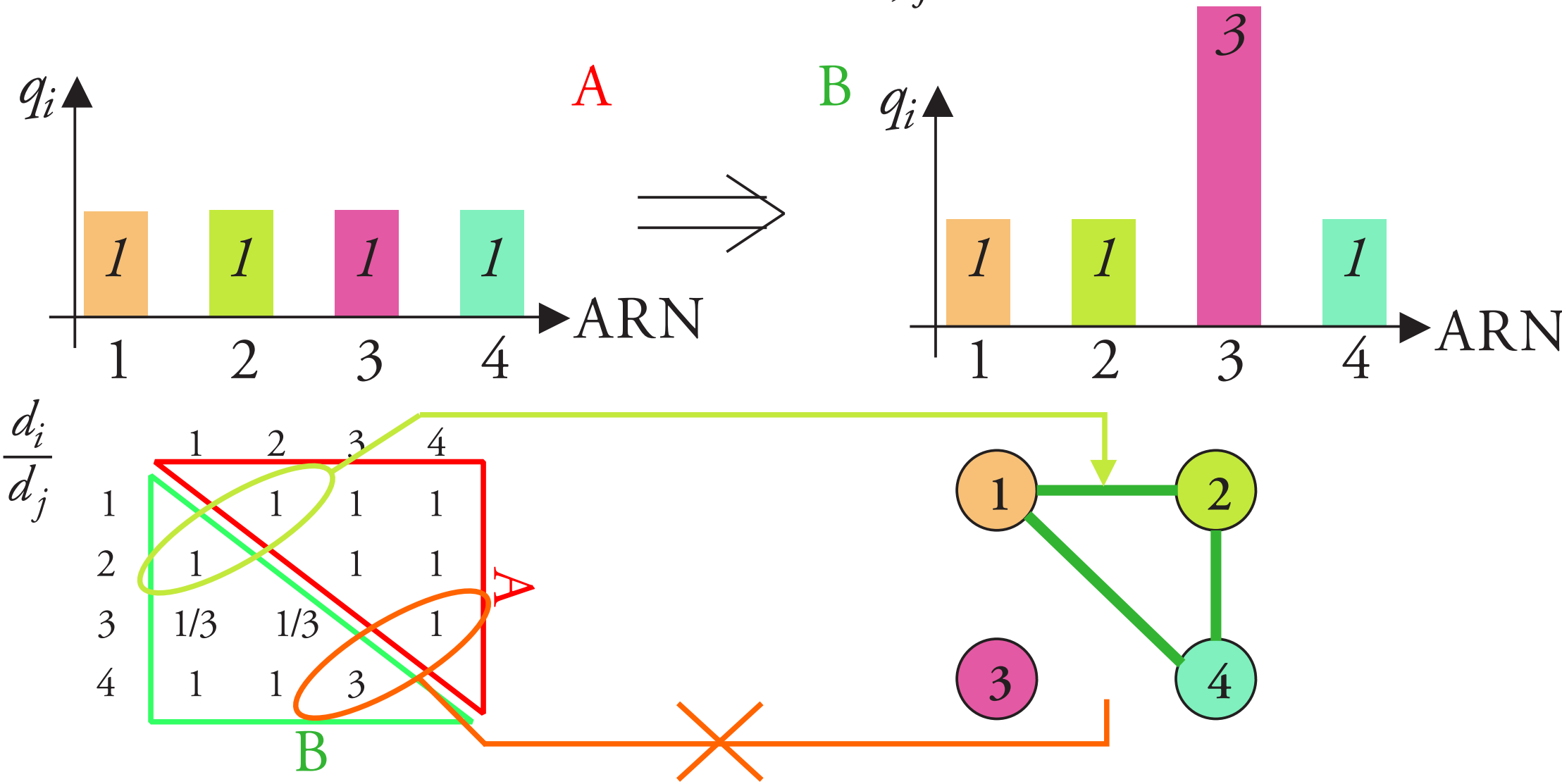
Le rapport est inchangé

- 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.

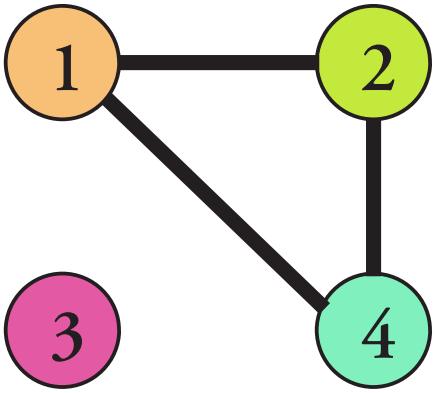
Construire un graphe des composants quantifiés

- 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…

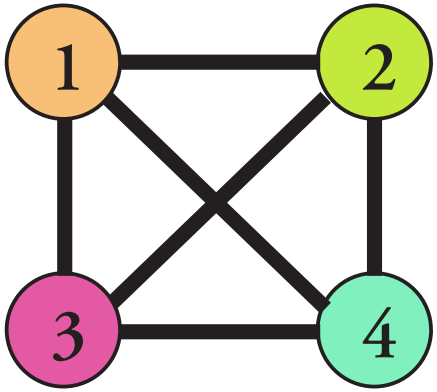
Approche constructive

- On part d'un graphe sans aucune arête
- On ajoute une arête entre deux constituants si l'on prouve qu'ils se comportent (suffisamment) pareil
  - ⇨ Utilisation de tests d'équivalence

Comment construire le graphe empirique ? ②

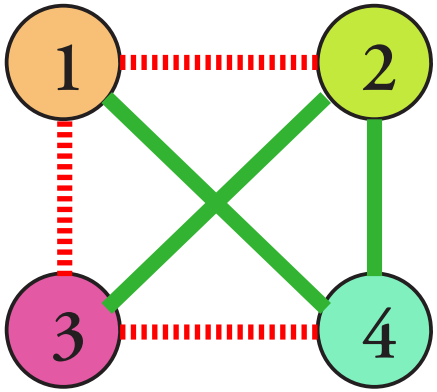
- Approche « destructive »
  - ⇨ On suppose le graphe complet (toutes les arêtes sont présentes)

Départ

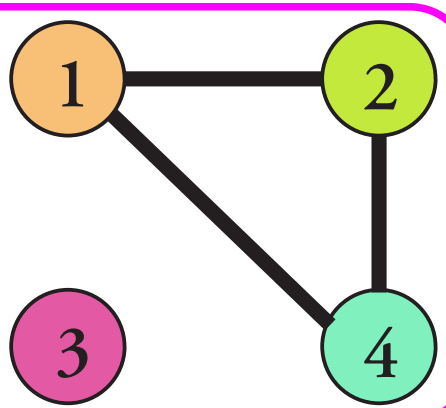


- On supprime l'arête entre deux constituants si l'on prouve qu'ils se comportent différemment
  - ⇨ Utilisation de tests de différence, « classiques »

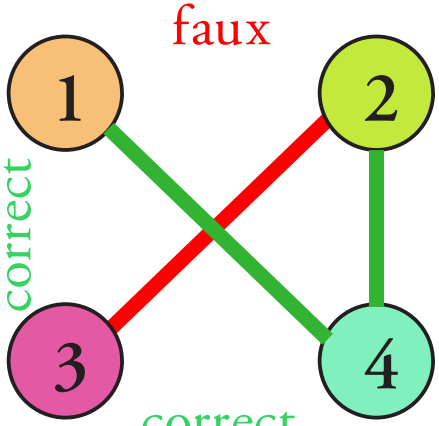
$p < \alpha$
$p > \alpha$



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 ?

Exemple 2 : Perfusion d'un traceur ①

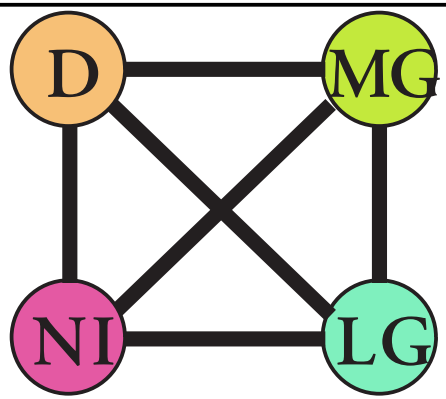
Données de Charles-Henry COTTART

- Modèle murin d'ischémie-reperfusion du foie
  - ⇨ Trois lobes ischémiés puis reperfusés (LMD, LMG, LLG)
  - ⇨ Deux lobes non-ischémiés (groupés en NI)
- Injection de micro-particules luminescentes (traceur)
- Quatre groupes de souris
  - ⇨ groupe témoin
  - ⇨ groupe ischémié, non-reperfusé
  - ⇨ groupes ischémiés et reperfusé, avec particules injectés après 20 s et 5 min de reperfusion

Exemple 2 : Perfusion d'un traceur ②

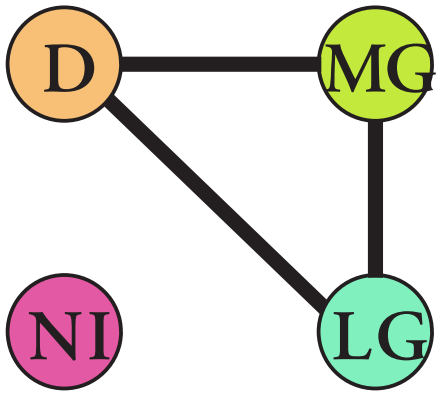
- Pour cet exemple, on se limite aux groupes « témoin » et « ischémié, non-reperfusé »
  - ⇨ Contrôle positif : l'ischémie modifie-t-elle la perfusion des lobes ? Normalement oui…
- Quantification par fluorescence des particules dans chaque lobe hépatique
  - ⇨ La quantité totale injectée est répartie entre ces lobes
  - ⇨ Données compositionnelles par nature !
- K = 4 « composants » — LMD, LMG, LLG et NI
  - ⇨ Quels graphes sont possibles ?

Exemple 2 : Les Principaux Graphes possibles



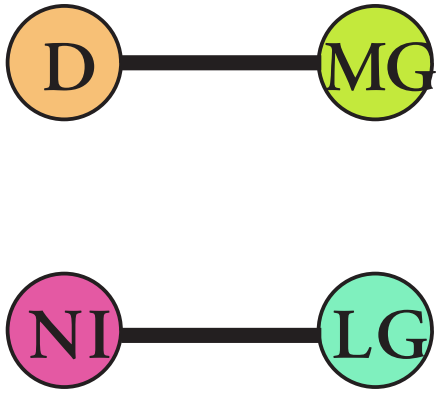
Aucune différence

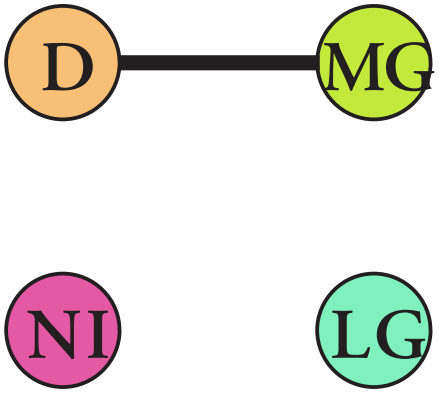
L'ischémie n'a pas eu lieu ou est totale
Expérience ratée...

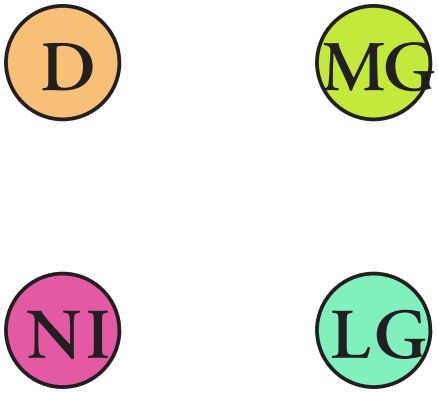


3 lobes similaires
1 lobe différent

Lobes NI : perfusion augmentée ou inchangée
Autres lobes : perfusion diminuée voire nulle
Expérience réussie...







Et variantes...

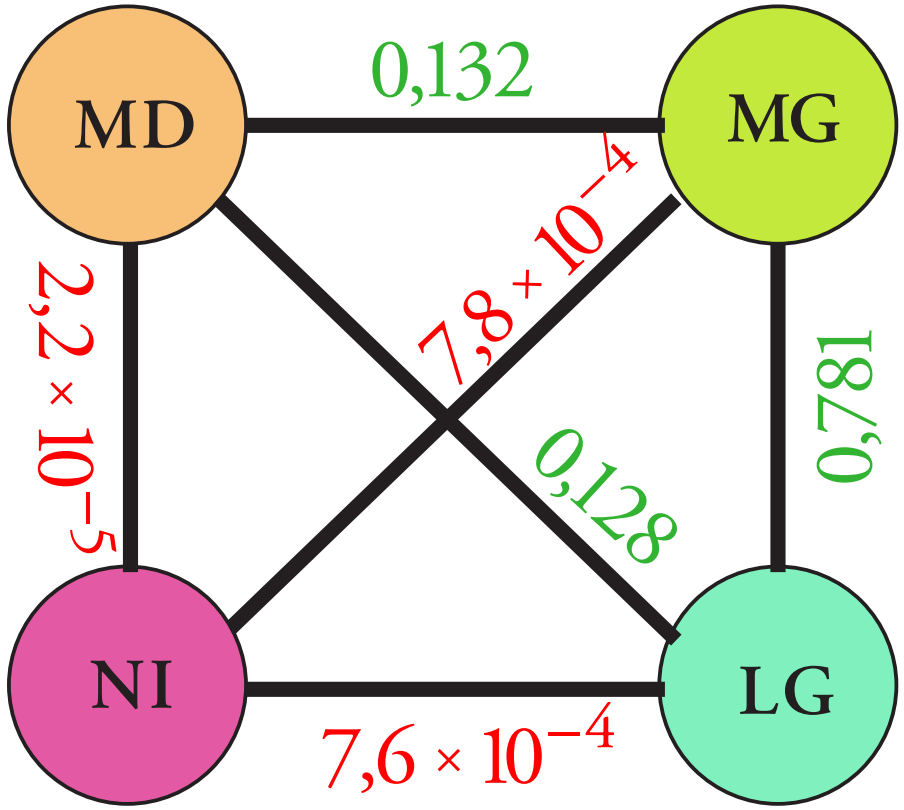
Ischémie partielle, différente entre les lobes

Exemple 2 : Le Graphe expérimental

- Effectif : $n = 5$ par groupe
- K = 4 donc 6 arêtes à tester
- Chaque arête correspond au rapport entre les deux lobes
- Pour chaque rapport, on teste s'il diffère entre les deux groupes

  ⇨ Test de Student sur le ln du rapport

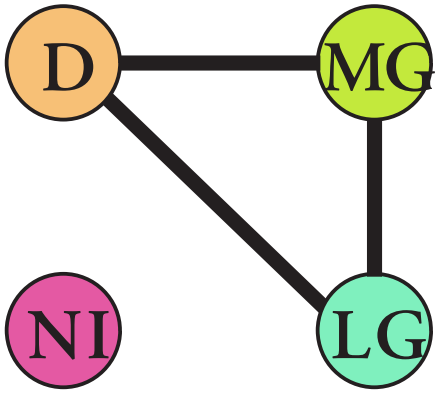
Résultats des tests



Résultats condensés : matrice de p

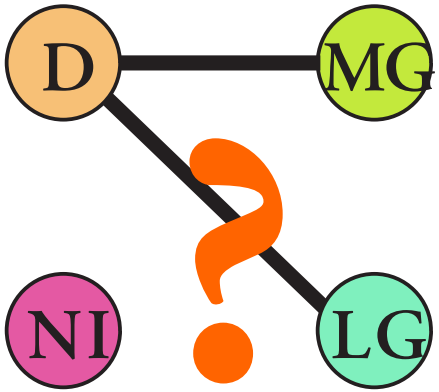
Exemple 2 : Conclusions... et questions !

- Graphe expérimental obtenu : celui ci-contre
  - ⇨ L'expérience semble avoir réussi



- Ici, cas facile : contrôle positif...
  - ⇨ La coupure ou non des arêtes est « facile »
  - ⇨ Le graphe obtenu est cohérent

- Sur des cas moins clair, comment faire ?
  - ⇨ Comment interpréter un graphe « impossible » ?
  - ⇨ À partir de quand couper une arête ?
  - ⇨ Quel est le risque de voir quelque chose qui n'existe pas ?



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

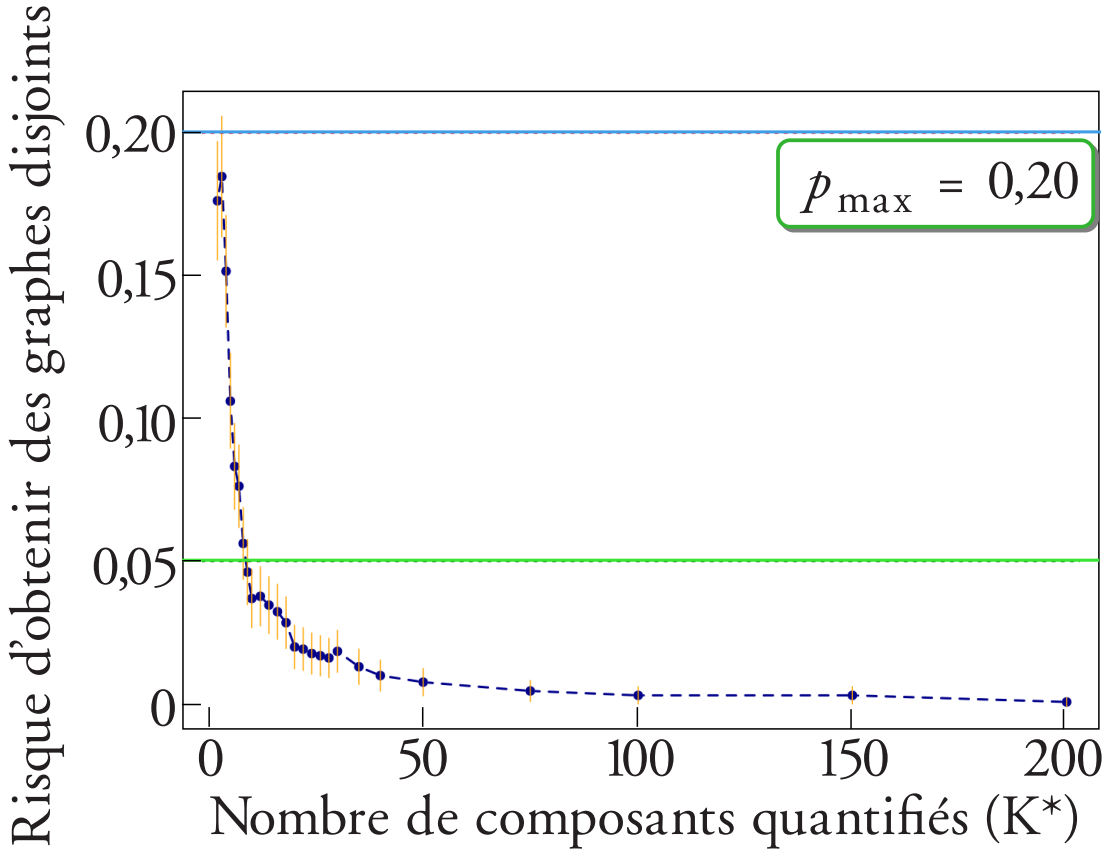
- Si les tests étaient indépendants, $\Pr(\text{TSM}|H_0) = \alpha = \alpha_0^{K^*-1}$
  - ⇨ α_0 doit être (bien) plus grand que α

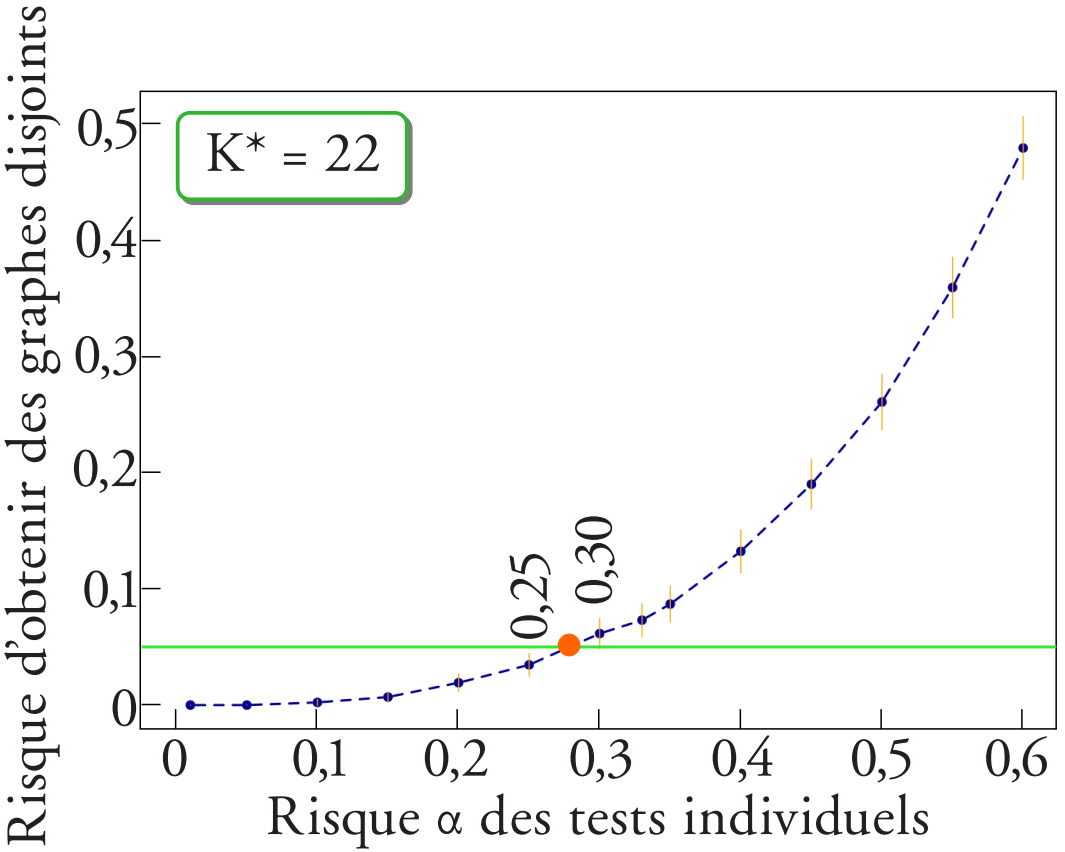
- Les tests ne sont **pas** indépendants. Et doit être vrai pour tous les nœuds…

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





Exemple 3 : 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

Exemple 3 : 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 la surface du champ observé
  - ⇨ 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

Exemple 3 : Composition tissulaire — approche naïve

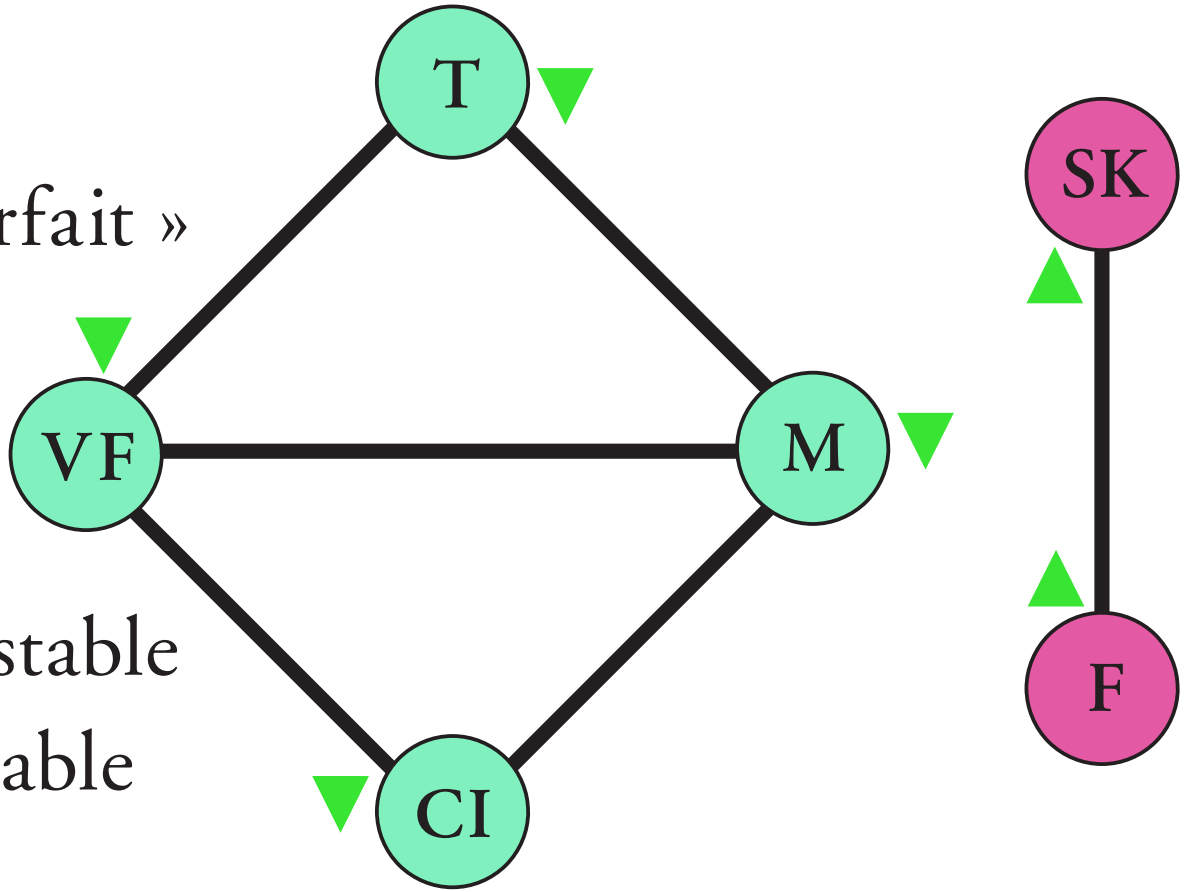
- Chaque tissu est testé isolément
  - ⇨ Néglige le côté compositionnel…
  - ⇨ 6 tests : travail avec $\alpha < 0{,}0083$ (Bonferroni)

Tissu	p	Sens
SK	0,0614	▲
F	0,0002	▲
T	0,0458	▼
VF	0,2066	▼
M	0,0952	▼
CI	0,0796	▼

- Conclusion : la fibrine (F) augmente

Exemple 3 : 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

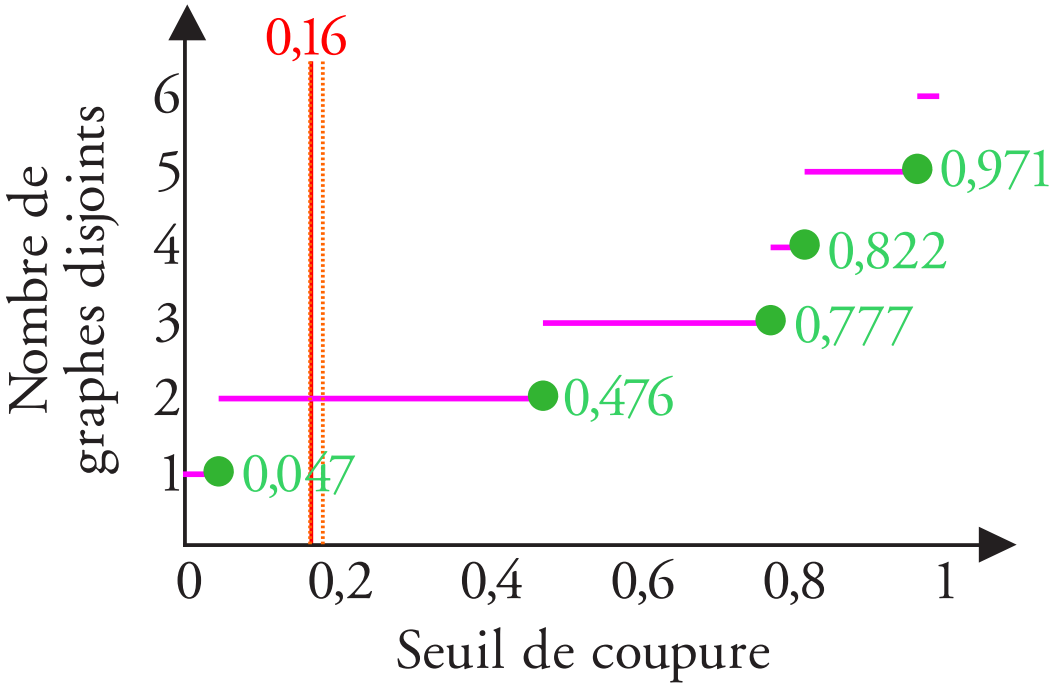


Quelle est la sensibilité du « découpage » au choix du seuil de coupure ?

Sensibilité au seuil de coupure (1)

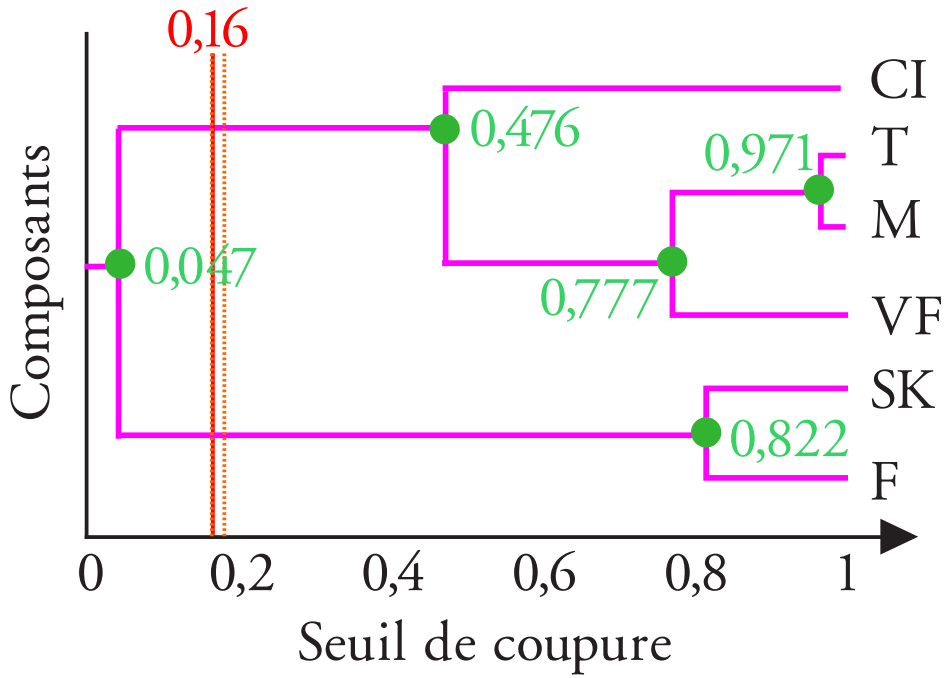
- Plus on augmente le seuil de coupure, plus on ôte d'arêtes
  - ⇨ Plus on a de chance d'avoir des graphes disjoints
  - ⇨ De combien faut-il bouger le seuil pour reconnecter les deux graphes ? ou les dissocier ?

- Il suffit d'essayer pour les p de la matrice, triés par ordre croissant
  - ⇨ Nombre de graphes disjoints en fonction du p seuil



Sensibilité au seuil de coupure ②

- À partir de ces coupures, on peut aussi proposer un arbre de classification des composants
  - ⇨ Arbre hiérarchique de séparation des groupes en fonction du seuil
  - ⇨ Plus deux composants se regroupent difficilement et plus ils se comportent différemment



Application en qRT-PCR

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] = $d_i = \lambda_i x_i = \lambda_i \mathrm{M} \dfrac{q_i}{\sum_{j=1}^{\mathrm{K}} q_j}$

Méthodes classiques ①

- Parmi les K^* A. R. N. quantifiés, K_R « gènes de référence »
  - ⇨ Gènes dont l'expression est postulée invariante
  - ⇨ Pour simplifier les notations : composants 1 à K_R

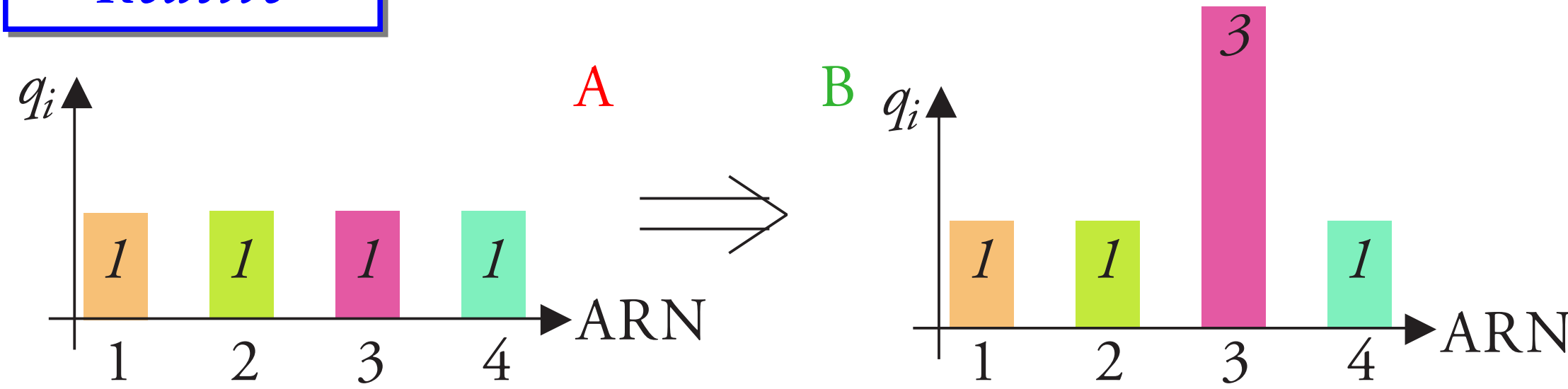
- Dans chaque échantillon, on calcule un facteur de normalisation, f_{norm}
  - ⇨ Moyenne arithmétique des d_i $(1 \leqslant i \leqslant K_R)$
  - ⇨ Moyenne géométrique des d_i (arithmétique des C_t)

- On étudie les rapports $r_{i,\,\text{norm}} = \dfrac{d_i}{f_{\text{norm}}}$ $\quad (K_R < i \leqslant K^*)$
  - ⇨ Si $K_R = 1$, méthode a. l. r. et « ΔC_t » !
  - ⇨ Il faut faire $(K^* - K_R)$ tests distincts

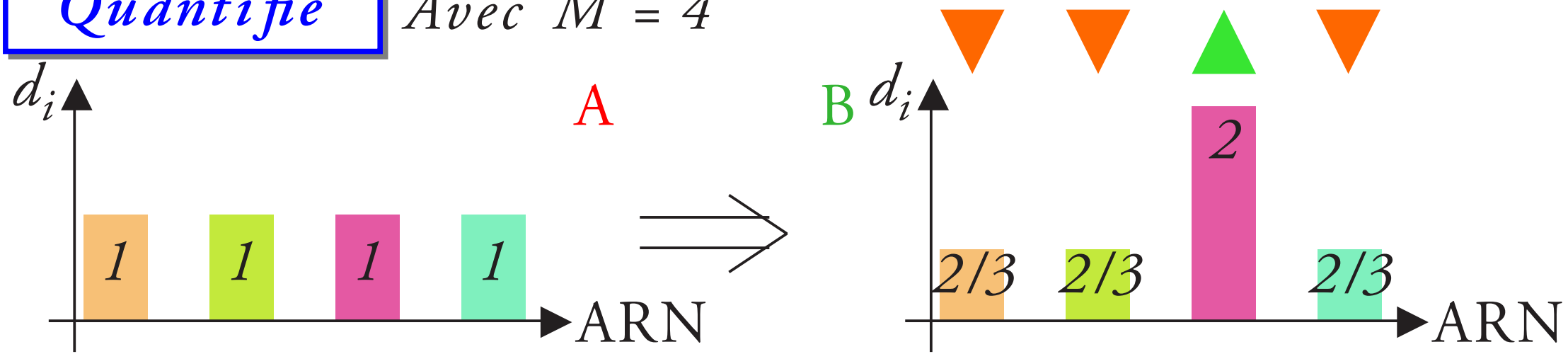
Conséquences sur les conclusions — Exemple ①

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

Réalité



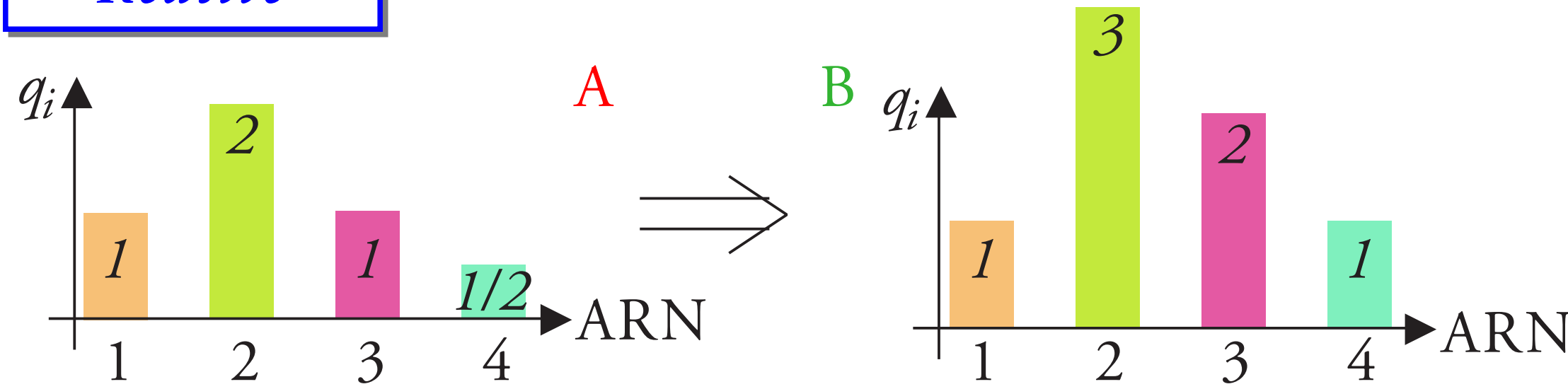
Quantifié Avec $M = 4$



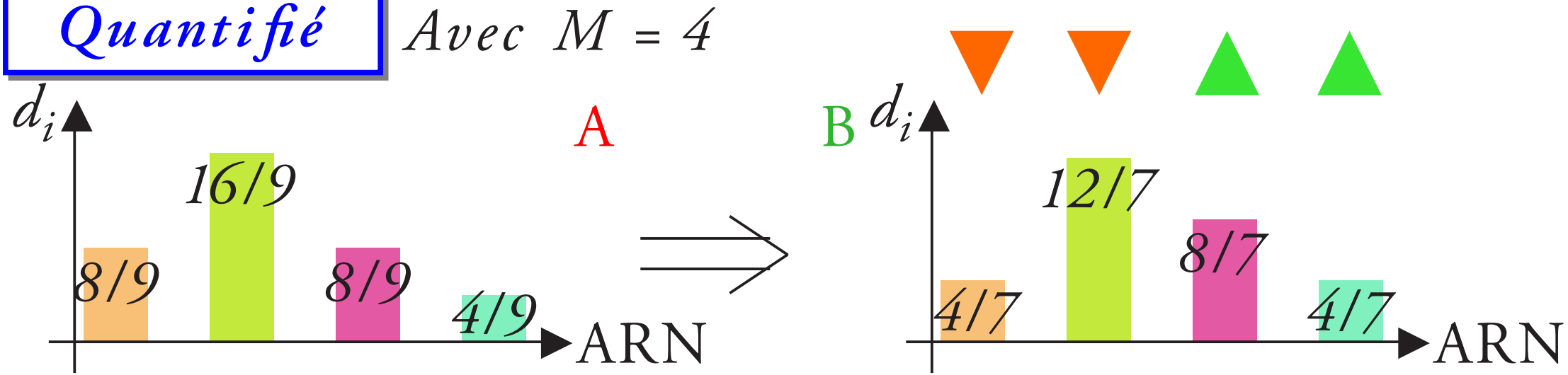
Conséquences sur les conclusions — Exemple ②

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

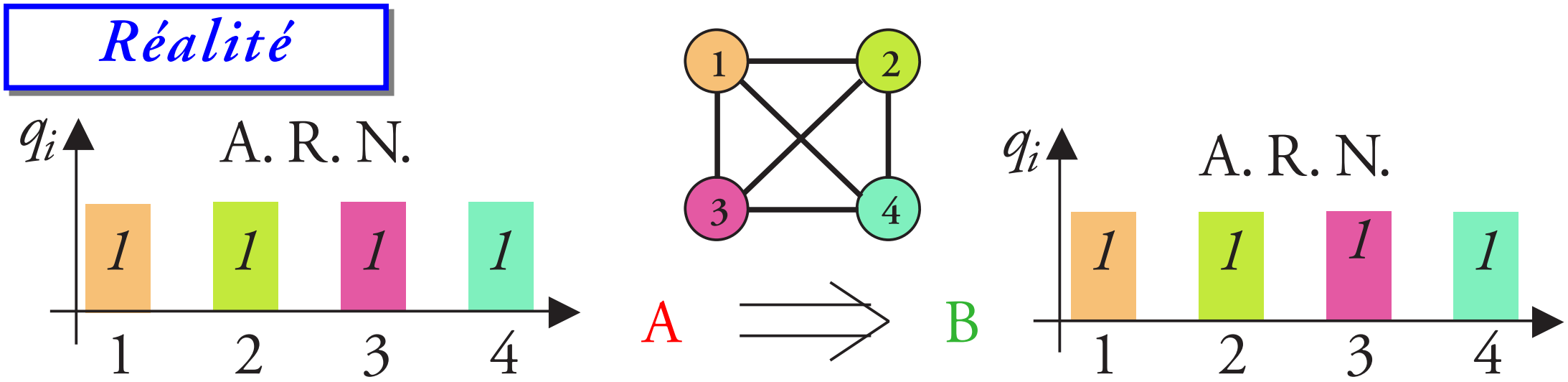
Réalité



Quantifié Avec $M = 4$



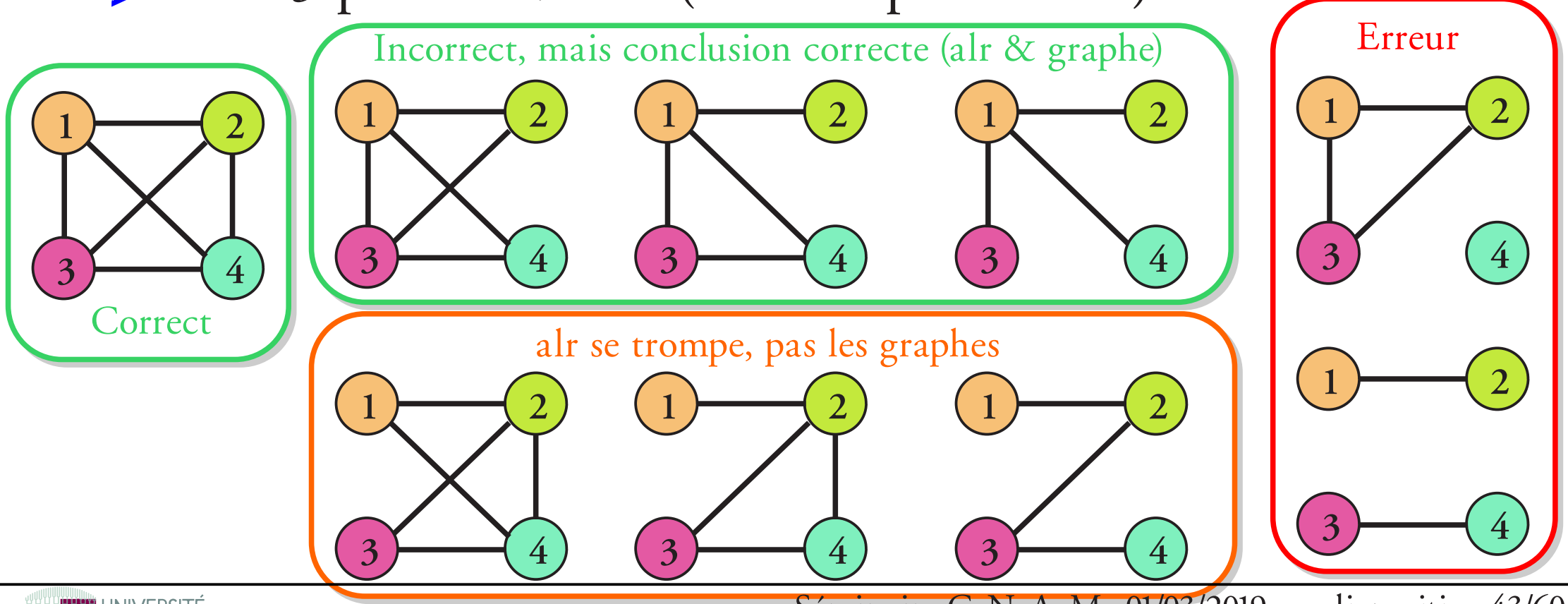
Étude sous H_0 — Principe



- Sous H_0, rien ne se passe entre A et B
  - ⇨ On ne doit obtenir qu'un seul graphe
  - ⇨ Méthode a. l. r. : on ne doit rien trouver
  - ⇨ Le pourcentage de simulations se trompant estime α
- Pour la méthode a. l. r., il faut une référence
  - ⇨ Arbitrairement, l'A. R. N. nº 1 sera la référence

Étude sous H_0 — conditions de simulation

- 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
  - ⇨ Pas du tout pour la méthode ΔC_t
  - ⇨ n = 3 par condition (raison : patience…)

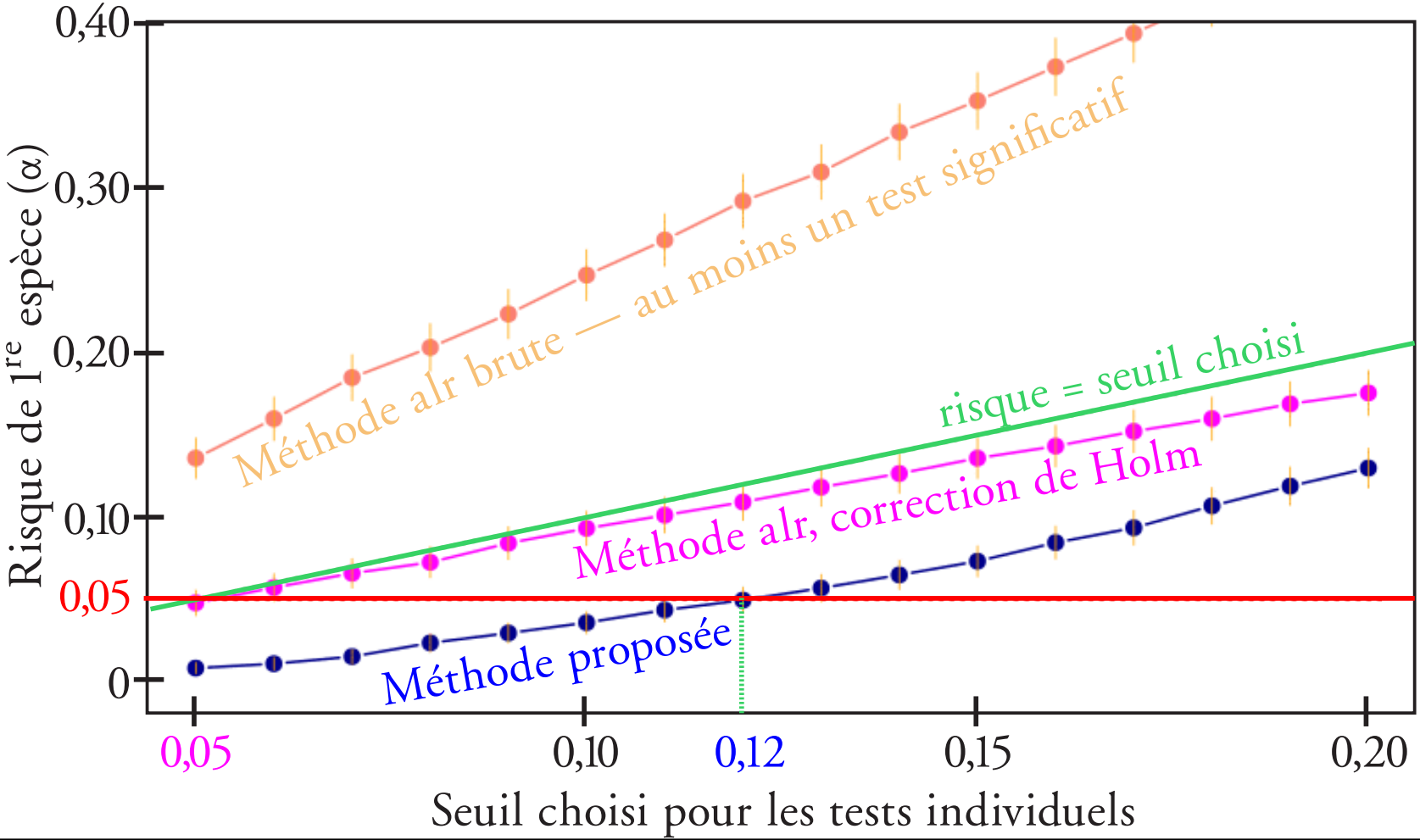


Étude sous H_0 — risque α

- La méthode alr « brute » ne contrôle pas α

  ⇨ Correction de multiplicité (ici, Holm) indispensable

- Méthode des graphes : coupure pour $p < 0{,}12$ ici



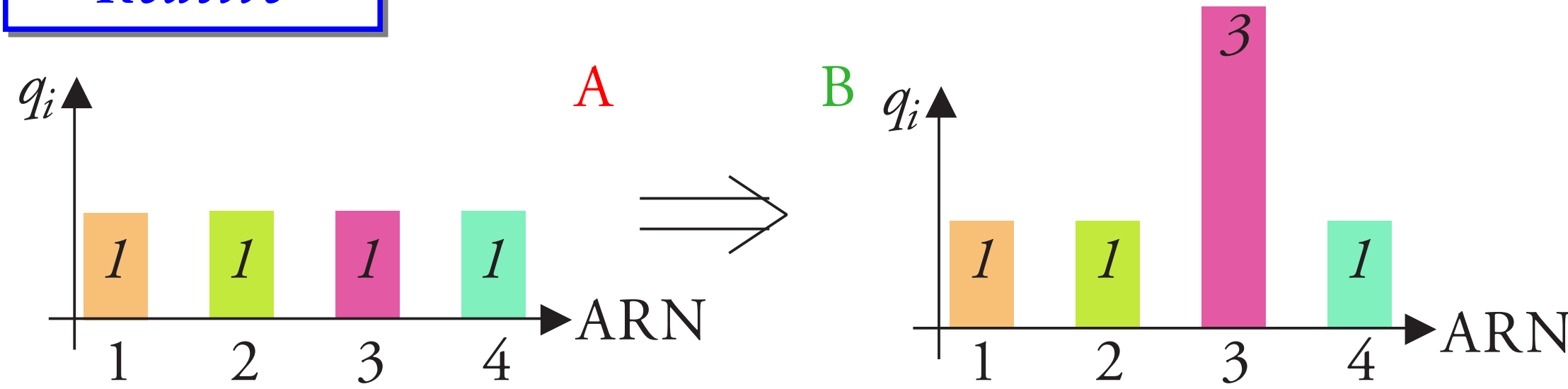
Et pour la puissance ?

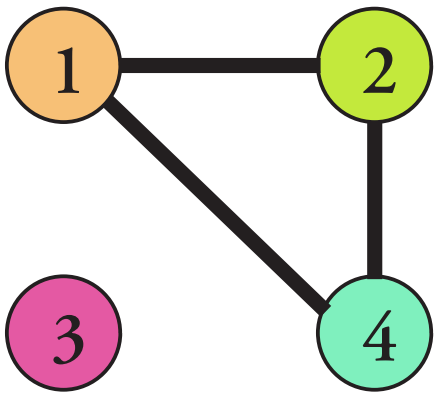
- Simulations précédentes : risque de détecter des composants se comportant différemment, quand tous se comportent de façon identique (« rejet de H_0 quand elle est vraie »)
- Comment se comporte la méthode quand certains composants se comportent réellement différemment ?
  - ⇨ Détecte-t-elle des graphes disjoints ? (*puissance*)
  - ⇨ Détecte-t-elle les bons groupes de composants ?
- Dépend de la puissance de chaque test individuel
- Difficile de comparer à la méthode classique alr
  - ⇨ Posent des questions légèrement différentes...

Étude de l'exemple 1 — conditions de simulation

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

Réalité





Graphe théorique

- Distribution log-normale ; CV = 20 %
- n = 3 par condition (puissance individuelle > 80 %)
- Méthode alr : référence = composant 1

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

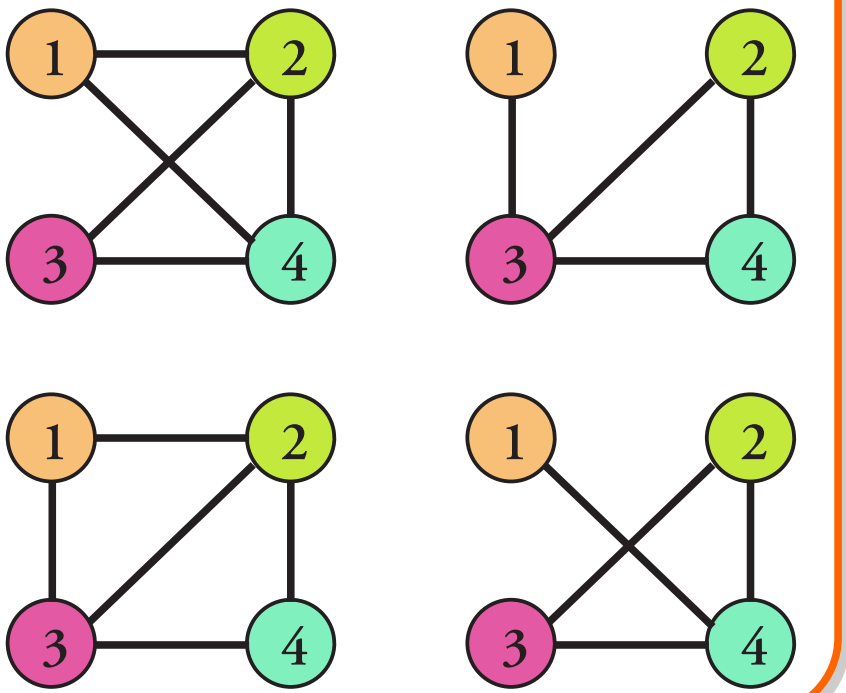
- On rejette H_0 pour n'importe quelle raison

  ⇨ Il se passe quelque chose (ce qui est observé… ou pas !)

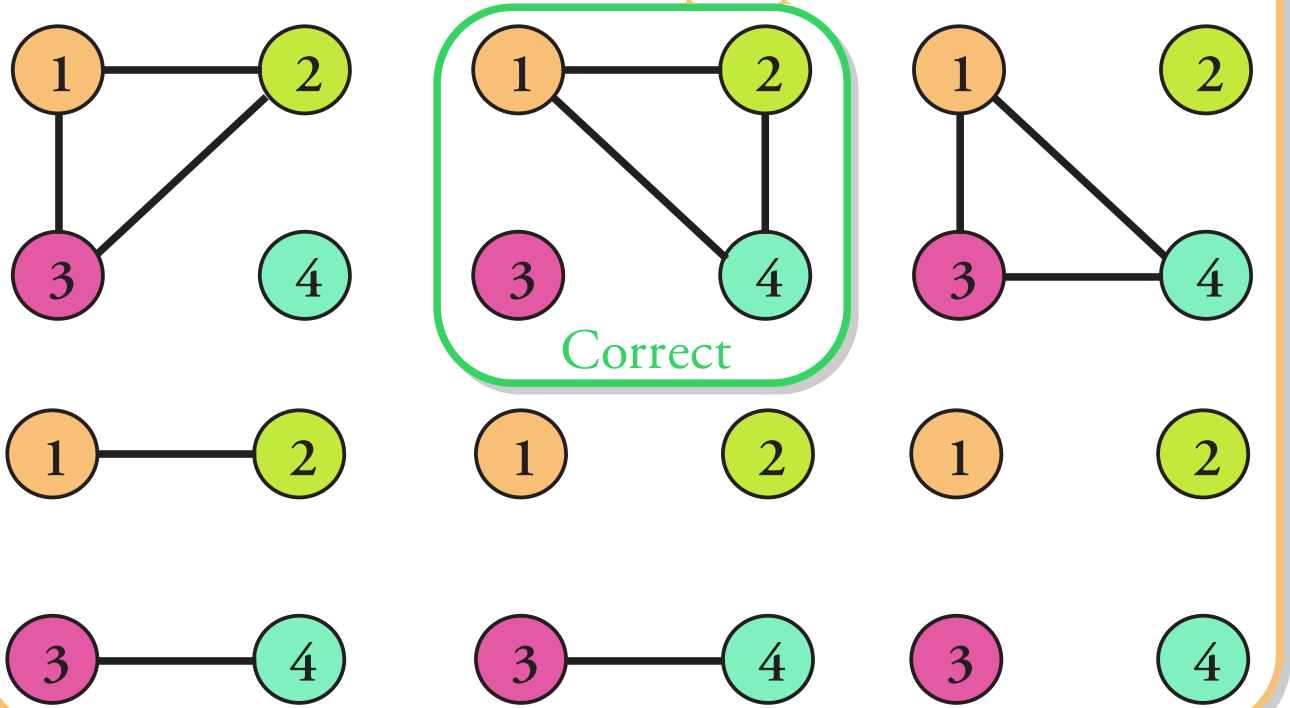
- Méthode alr : composants 2 ou 3 ou 4 détectés

  Méthode proposée : graphe disjoint, quel qu'il soit

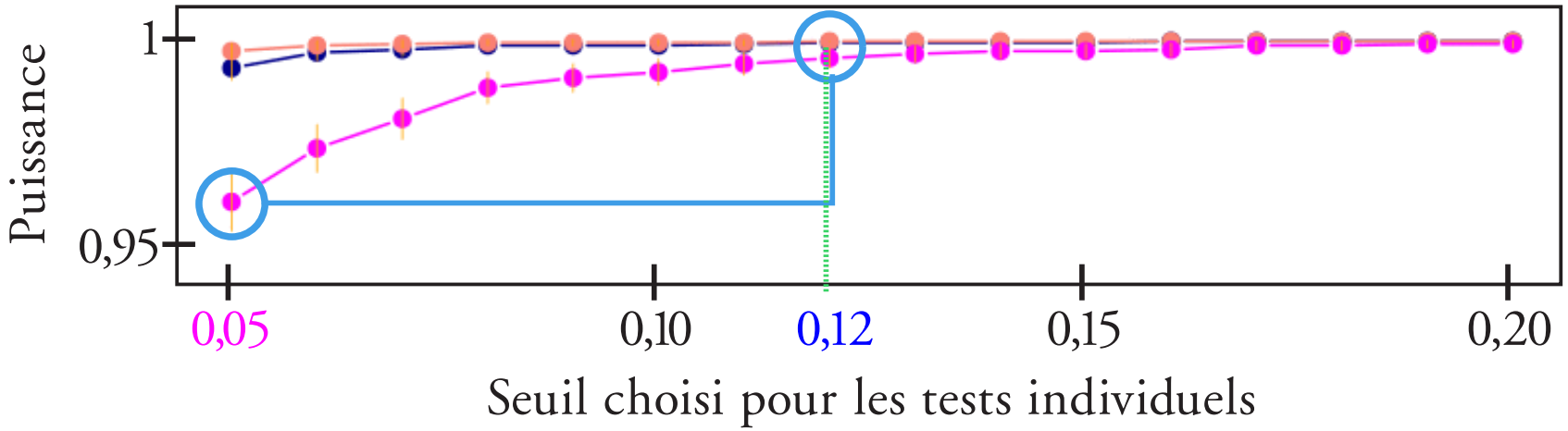
alr seule détecte quelque chose



Les deux détectent quelque chose

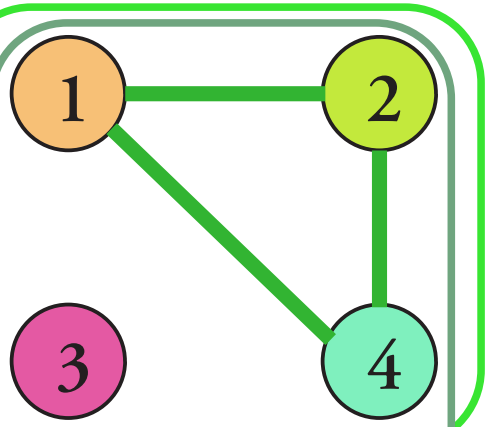


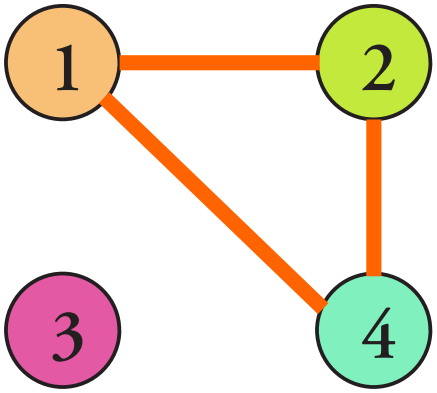
Étude de l'exemple 1 — étude de la puissance totale



- En pratique les deux méthodes sont aussi puissantes
  - ⇨ Mais la correction de multiplicité en alr a un prix !
- Mais est-ce qu'elles détectent bien ce qu'il faut ?
  - ⇨ alr : le composant 3 (et seulement lui) est significatif
  - ⇨ graphes : deux graphes, un avec le nœud 3 et un second avec les trois autres nœuds

Exemple 1 — trouve-t-on le bon composant ?

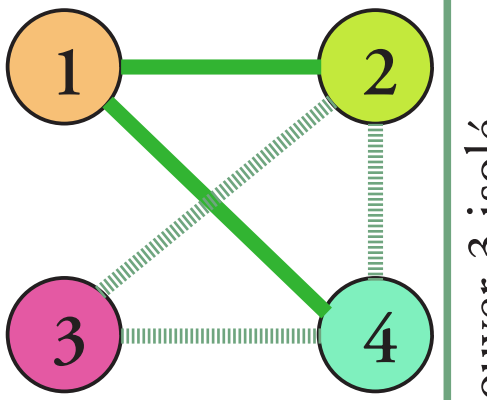




Méthode proposée

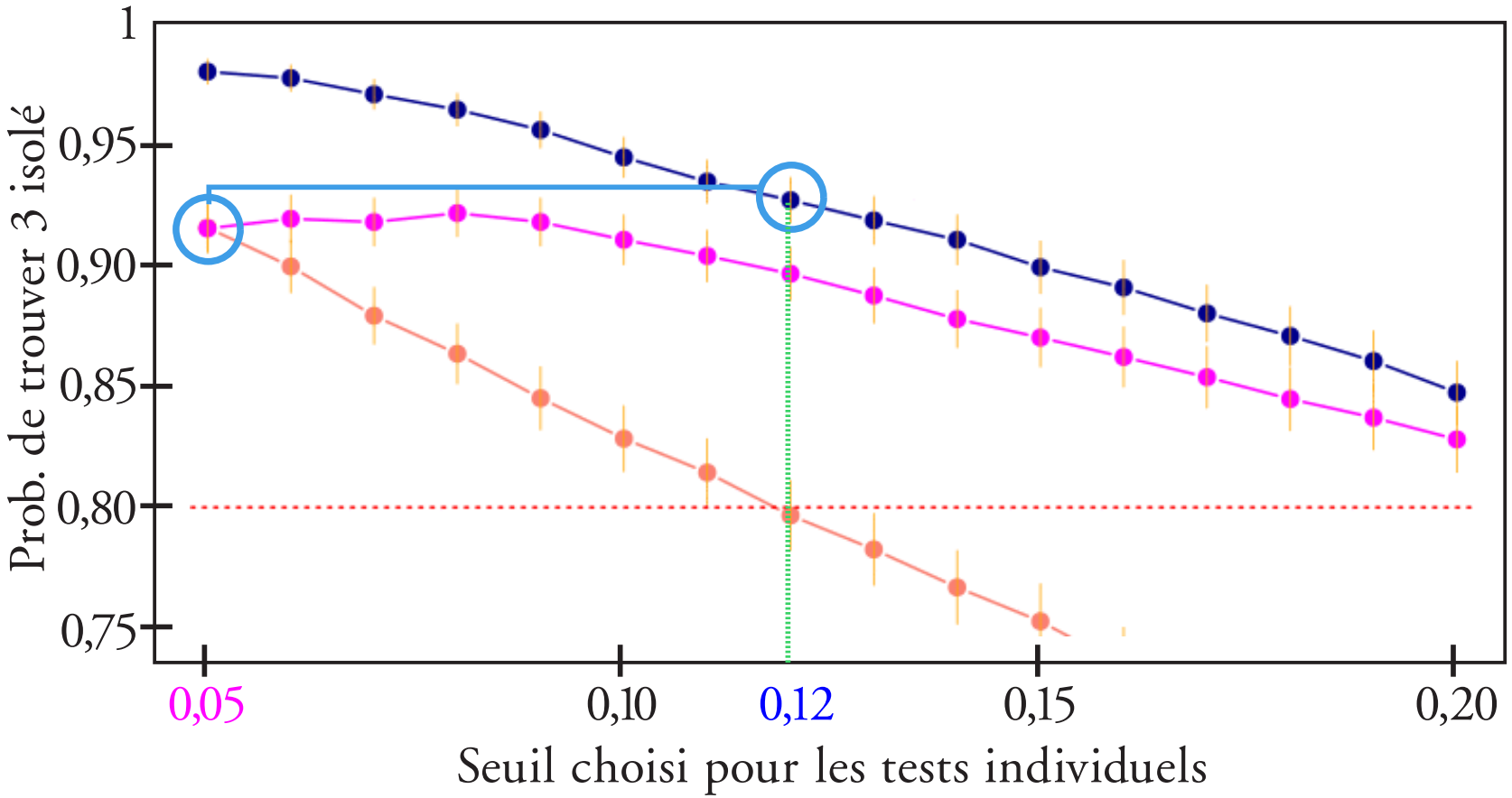
Une seule de ces arêtes peut manquer, au plus, pour conclure correctement

Soit 4 graphes acceptables (sur 2^6 possibles)



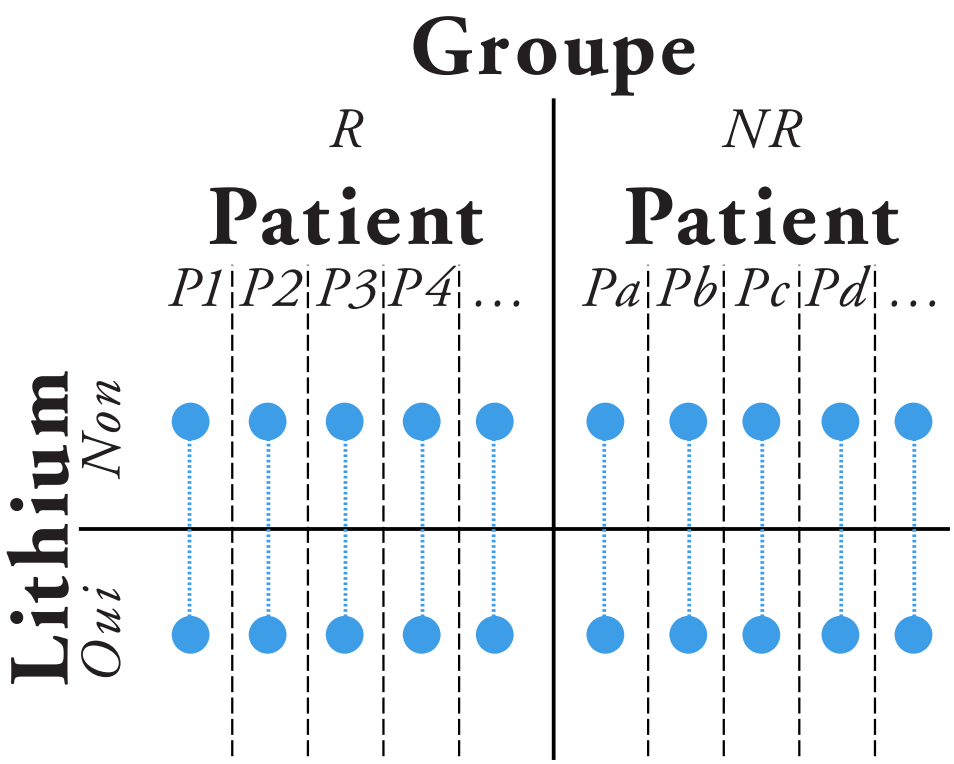
Méthode alr

Ces arêtes peuvent, ou non, exister — la conclusion est identique



Application en qRT-PCR: un exemple ①

- 2 groupes de patients bipolaires (répond [R, $n = 19$] ou non [NR, $n = 19$] au lithium)
- Leurs lymphocytes sont cultivés avec et sans lithium
- 17 candidats + 2 références



Geoffroy et coll., 2017

Modèle

$$\ln r_{i,j} = \mu_0 + \underline{U_P} + \underline{\delta_R}\,\mathbf{1}_R + \underline{\delta_{Li}}\,\mathbf{1}_{Li} + \underline{\delta_I}\,\mathbf{1}_R\,\mathbf{1}_{Li} + \varepsilon$$

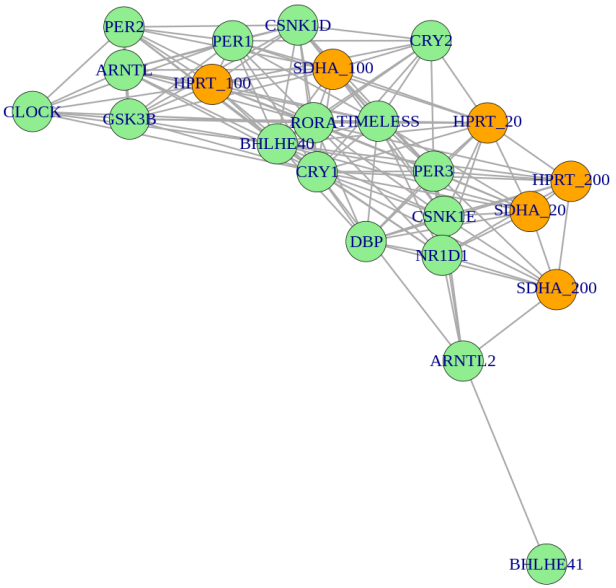
- U_P : *Effet patient*
- δ_R : *Différence basale*
- δ_{Li} : *Effet du lithium chez les patients NR*
- δ_I : *Effet additionnel du lithium chez les patients R*

Application en qRT-PCR : un exemple ②

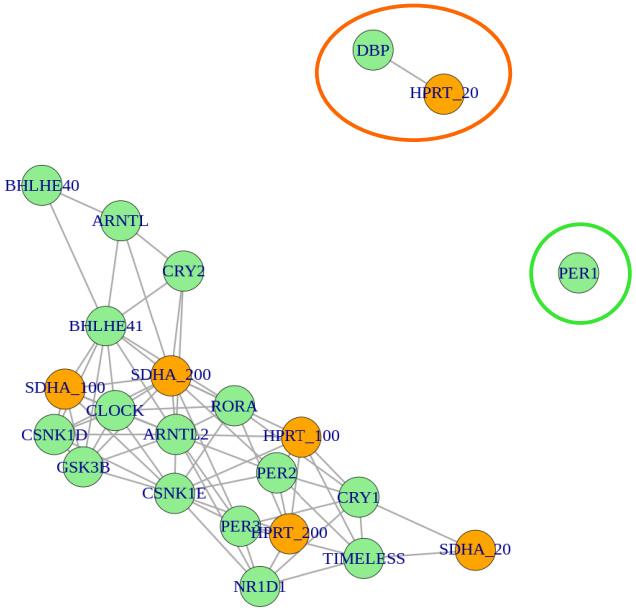
- Modèle linéaire à effets mixtes, avec lme4 (R)
  - ⇨ Coefficients testés par test de Wald asymptotique

- Critère : sous-graphes disjoints
- K* = 23 nœuds ⇨ α_0 = 0,25 pour avoir α < 0.05

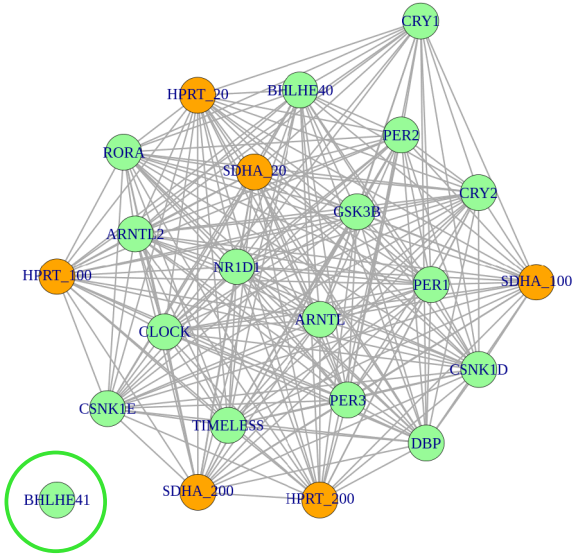
Comparaison R vs NR, basal



Effet de Li⁺ chez les NR



Effet différentiel de Li⁺



Application en 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, $[C_i] = \varepsilon_i \, q_i$
  - ⇨ 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

Exemple d'application ①

Étude de la maladie du greffon contre l'hôte

- 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

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

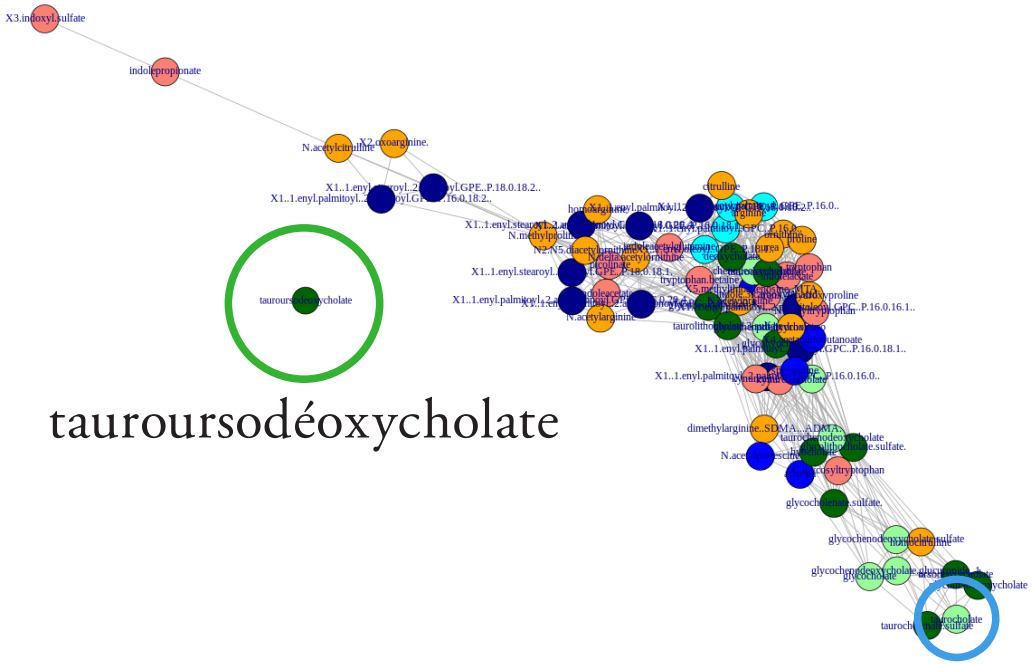
Exemple d'application ② — Résultats

- Analyse sur 70 métabolites (voies sélectionnées)

⇨ Seuil de coupure : $p < 0{,}35$

Cohorte 1

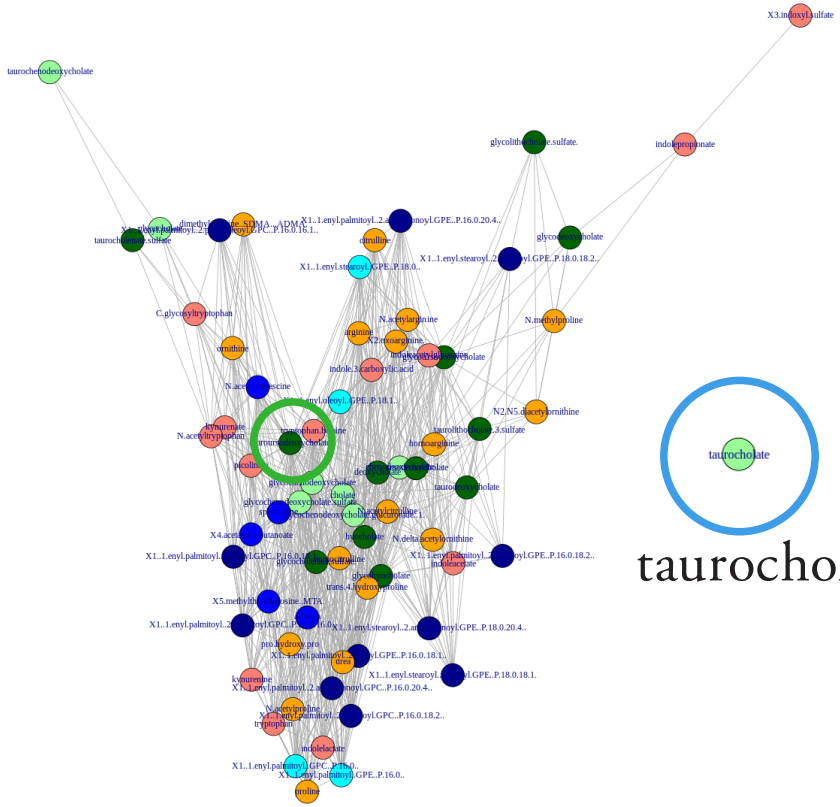
12 *vs* 31 patients



tauroursodéoxycholate

Cohorte 2

26 *vs* 30 patients



taurocholate

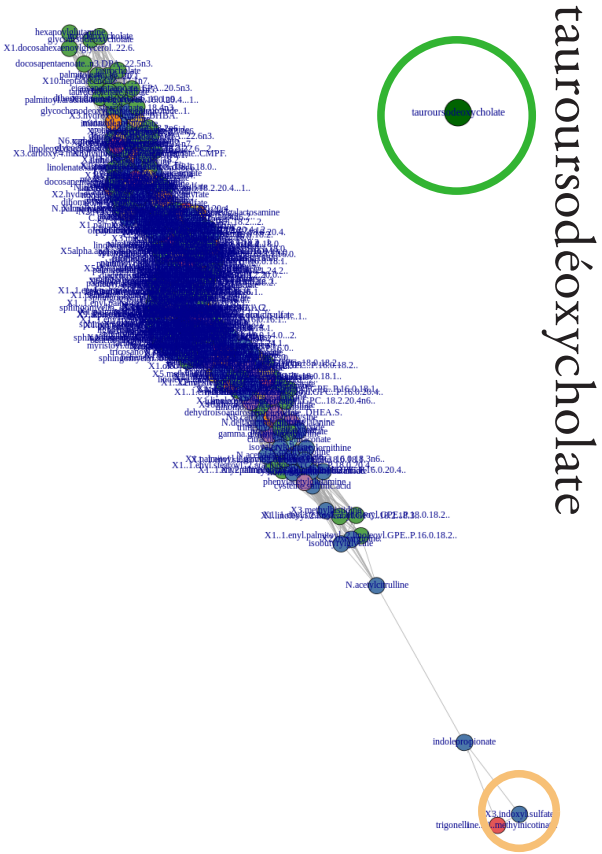
Exemple d'application ③ — Résultats

- Analyse sur 491 métabolites (communs aux deux cohortes)

⇨ Seuil de coupure : $p < 0{,}44$

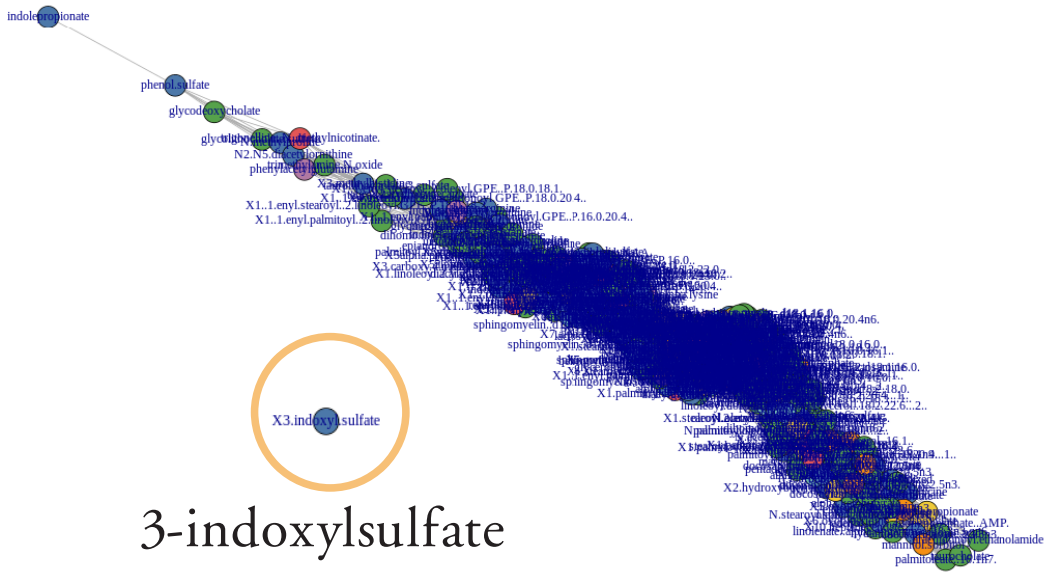
Cohorte 1

12 *vs* 31 patients



Cohorte 2

26 *vs* 30 patients



3-indoxylsulfate

Quelques limitations de la méthode...

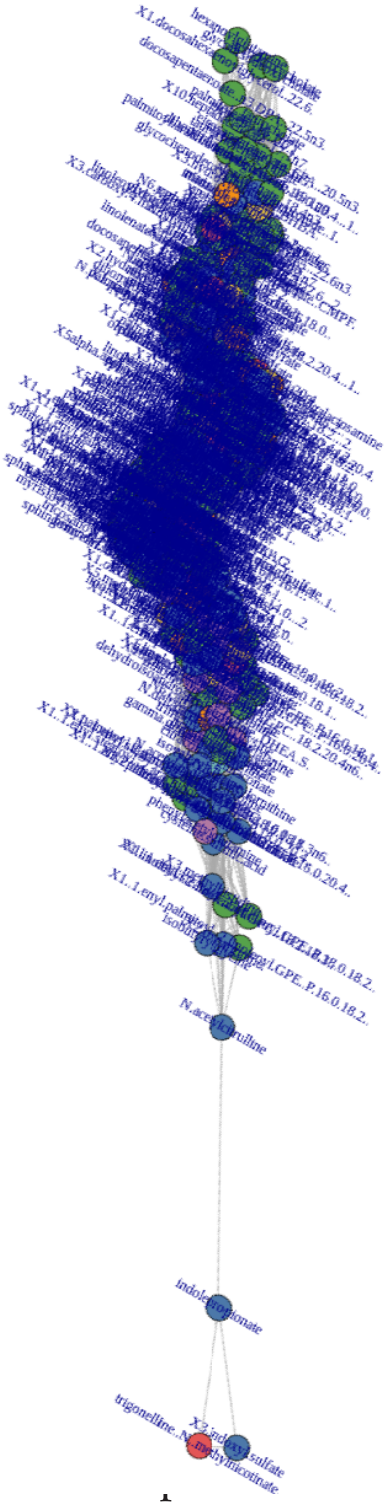
- Nombre de tests augmente avec K* comme K^{*2}
  - ⇨ Temps et mémoire nécessaires augmentent « vite »
  - ⇨ Temps de calcul pour K = 800 : 15 minutes
  - ⇨ Temps de 1000 simulations, 48 cœurs : 5 heures
  - ⇨ Peut être problématique en RNAseq ($K^* \approx 2 \times 10^4$)
- Taille du graphe augmente 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 !
- RNASeq : insensible aux profondeurs de séquençage…
- Applicable à tout plan expérimental

Perspectives

- 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



Remerciements

- Charles-Henry Cottart — données hépatiques (traceur)
- Anne-Gaelle Cordier — données d'imagerie (placentas)
- Cindie Courtin, Calypso Nepost, Pierre-Alexis Geoffroy — qRT-PCR
- David Michonneau, Gérard Socié — données urinaires
- Bruno Saubaméa — données de RNA-Seq
- Étudiants de l'ENSAI
- Sylvie Chevret, Yves Rozenholc, Chantal Guihenneuc

Merci de votre attention !

Bibliographie

Logiciel

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

Articles

Théorie

- E. Curis et coll., *Bioinformatics*, janvier 2019

Application

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