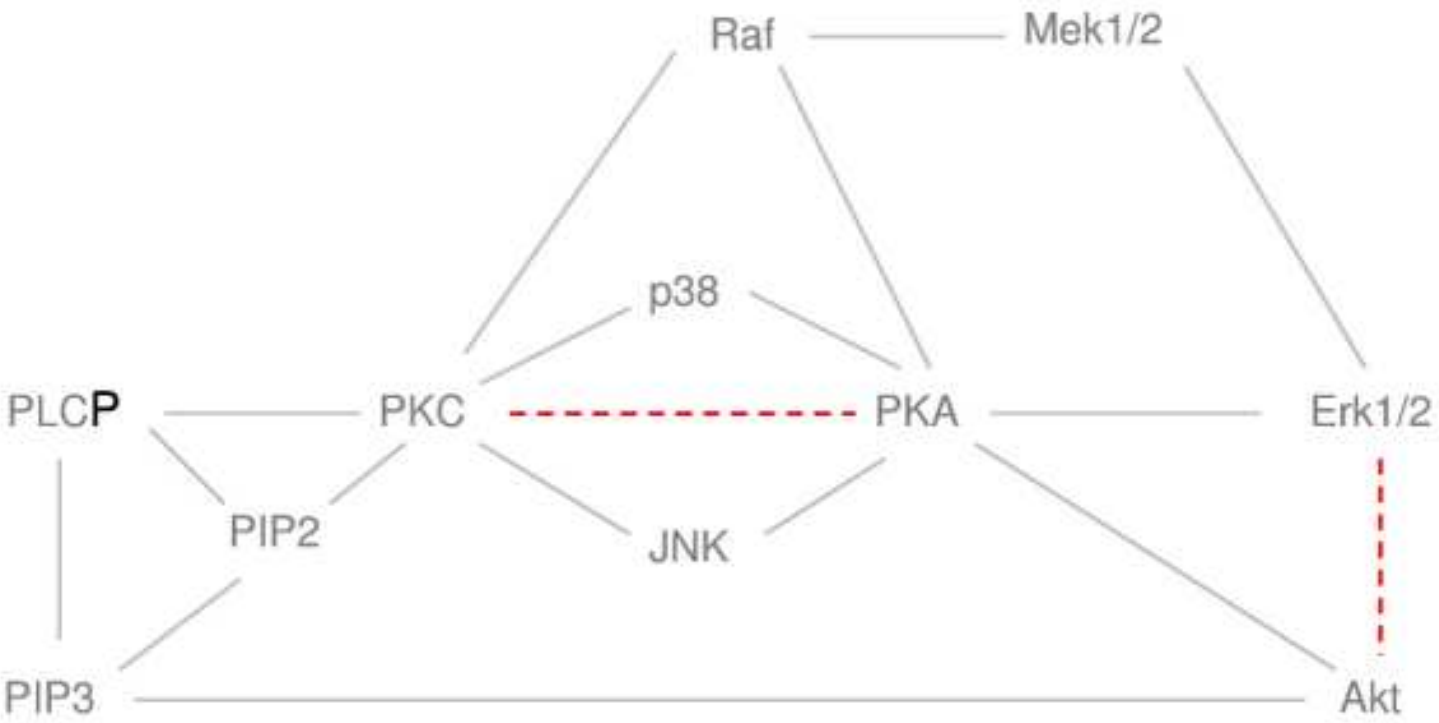
Test de validation de graphe

Fanny Villers, Nicolas Verzelen

INRA, unité MIA, Jouy-en-Josas
Université Paris-Sud

Rennes, 28 aout 2008

Raf pathway. (Sachs et al., 2005)



Les données

- expressions de gènes obtenues à partir de microarrays : $n \ll p$
- quantités de protéines obtenues à partir d'E2D : $n \ll p$
- données de cytométrie de flux : $n \gg p$

Objectifs différents :

- **Estimation de graphe**

  BUT : Estimer les relations entre les gènes ou protéines à partir de données

  références :
  - Schäfer et Strimmer (2004)
  - Wille et Bühlmann (2006)
  - Meinshausen et Bühlmann (2006)

Objectifs différents :

- **Estimation de graphe**

  BUT : Estimer les relations entre les gènes ou protéines à partir de données

  références :
  - Schäfer et Strimmer (2004)
  - Wille et Bühlmann (2006)
  - Meinshausen et Bühlmann (2006)

- **Validation de graphe**

  Connaissances expérimentales $\rightsquigarrow$ graphe minimal G

  BUT : Test de validation de ce graphe minimal :
  Tester à partir de données qu'on n'a pas oublié d'interaction.

Objectifs différents :

- **Estimation de graphe**

  BUT : Estimer les relations entre les gènes ou protéines à partir de données

  références :
  - Schäfer et Strimmer (2004)
  - Wille et Bühlmann (2006)
  - Meinshausen et Bühlmann (2006)

- **Validation de graphe**

  Connaissances expérimentales $\rightsquigarrow$ graphe minimal G

  BUT : Test de validation de ce graphe minimal :
  Tester à partir de données qu'on n'a pas oublié d'interaction.

On veut s'autoriser $p >> n$

Plan

Modèles Graphiques Gaussiens

$G = (\Gamma, \mathcal{A})$

- sommets $\Gamma = \{1, \ldots, p\}$: gènes, protéines, ...

- arêtes $\mathcal{A}$: liens fonctionnels directs entre les gènes

Modèles Graphiques Gaussiens

$G = (\Gamma, \mathcal{A})$

- sommets $\Gamma = \{1, \ldots, p\}$: gènes, protéines, ...
  $\equiv$ **variables aléatoires:** $X_1, \ldots, X_p$
  où X_a : expression du gène a

$$X = (X_1, \ldots, X_p) \sim \mathcal{N}_p(0, \Sigma)$$

- arêtes $\mathcal{A}$: liens fonctionnels directs entre les gènes

Modèles Graphiques Gaussiens

$G = (\Gamma, \mathcal{A})$

- sommets $\Gamma = \{1, \ldots, p\}$: gènes, protéines, ...
  $\equiv$ **variables aléatoires:** $X_1, \ldots, X_p$
  où X_a : expression du gène a

$$X = (X_1, \ldots, X_p) \sim \mathcal{N}_p(0, \Sigma)$$

- arêtes $\mathcal{A}$: liens fonctionnels directs entre les gènes
  Absence d'arête $\equiv$ indépendance conditionnelle :

Modèles Graphiques Gaussiens

$G = (\Gamma, \mathcal{A})$

- sommets $\Gamma = \{1, \ldots, p\}$: gènes, protéines, ...
  $\equiv$ **variables aléatoires:** $X_1, \ldots, X_p$
  où X_a : expression du gène a

$$X = (X_1, \ldots, X_p) \sim \mathcal{N}_p(0, \Sigma)$$

- arêtes $\mathcal{A}$: liens fonctionnels directs entre les gènes
  Absence d'arête $\equiv$ indépendance conditionnelle :

$$(a, b) \in \mathcal{A} \iff X_a \text{ dépendante de } X_b \mid X_{-(a,b)}$$

$$\iff \mathrm{Cov}\left(X_a, X_b | X_{-(a,b)}\right) \neq 0$$

$$\iff K_{ab} \neq 0$$

où $K = \Sigma^{-1}$: matrice de précision

Données :

$$X^i = (X_1^i, \ldots, X_p^i), \quad i = 1, \ldots, n$$

sont n observations indépendantes de même loi $\mathcal{N}_p(0, \Sigma)$

Données :

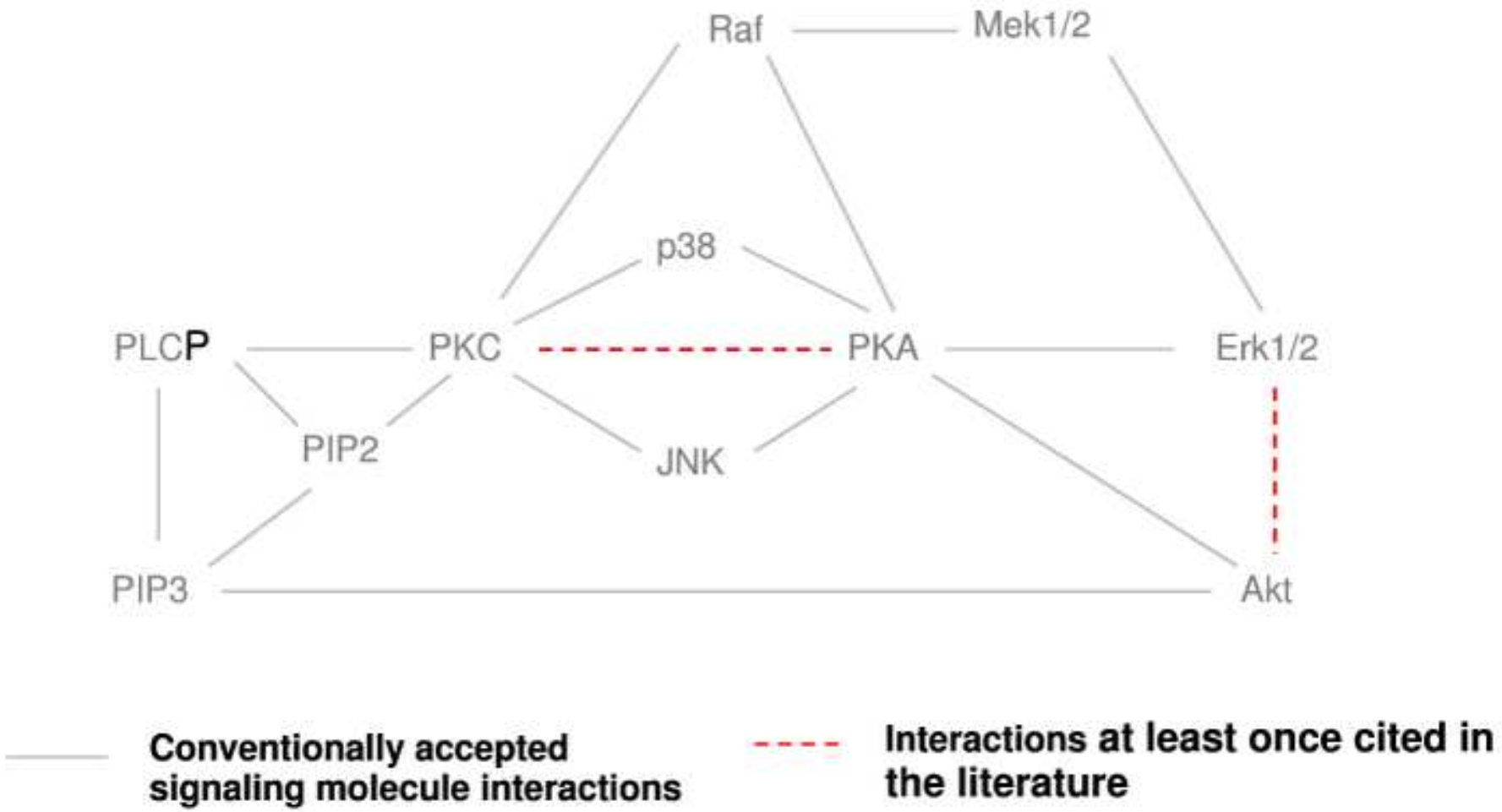
$$X^i = (X_1^i, \ldots, X_p^i), \quad i = 1, \ldots, n$$

sont n observations indépendantes de même loi $\mathcal{N}_p(0, \Sigma)$

Objectif : Test de validation de graphe

BUT : Valider à partir des n observations un graphe minimal obtenu à partir de connaissances préalables.

Raf pathway. (Sachs et al., 2005)



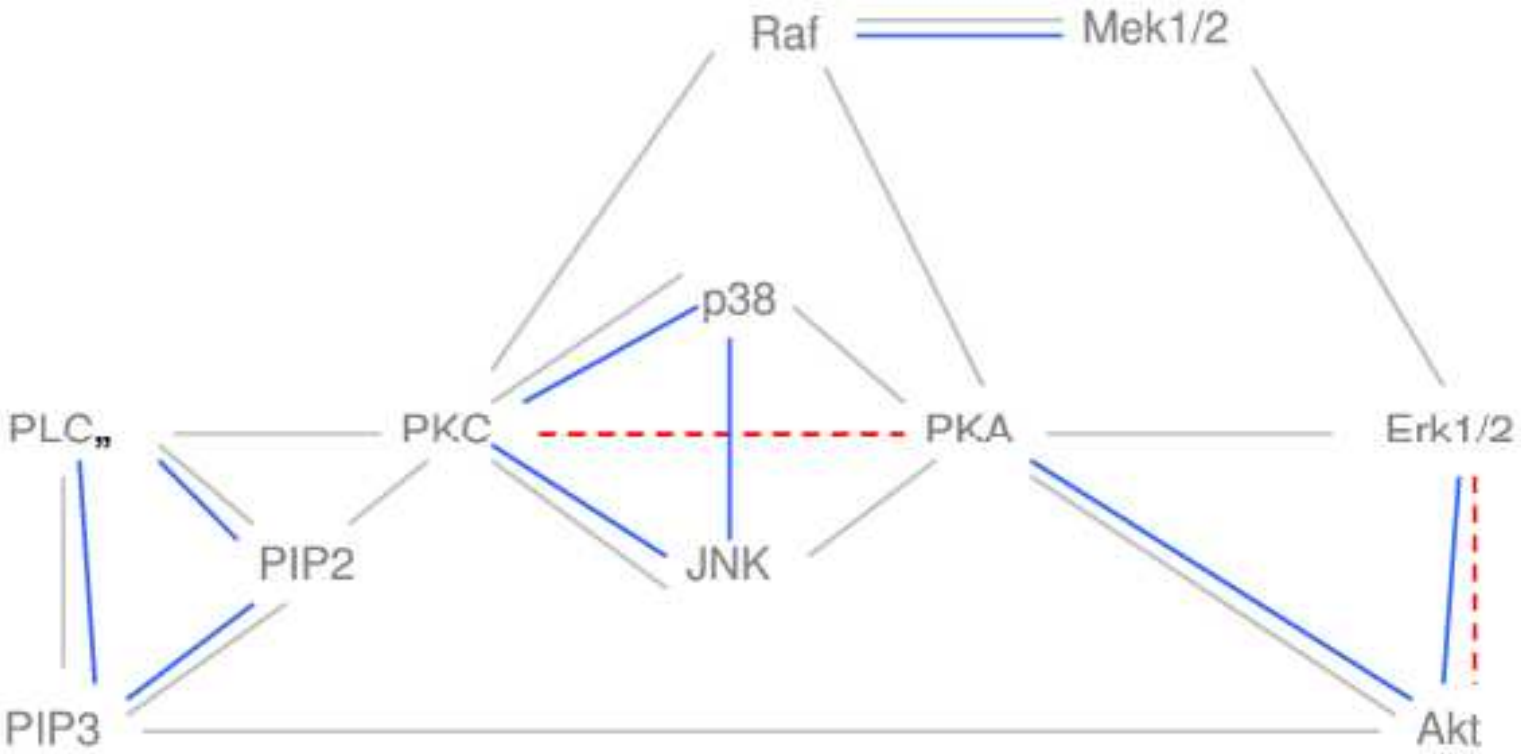
Resultat d'une méthode d'estimation à partir des données de Sachs

données : Sachs et al., 2005

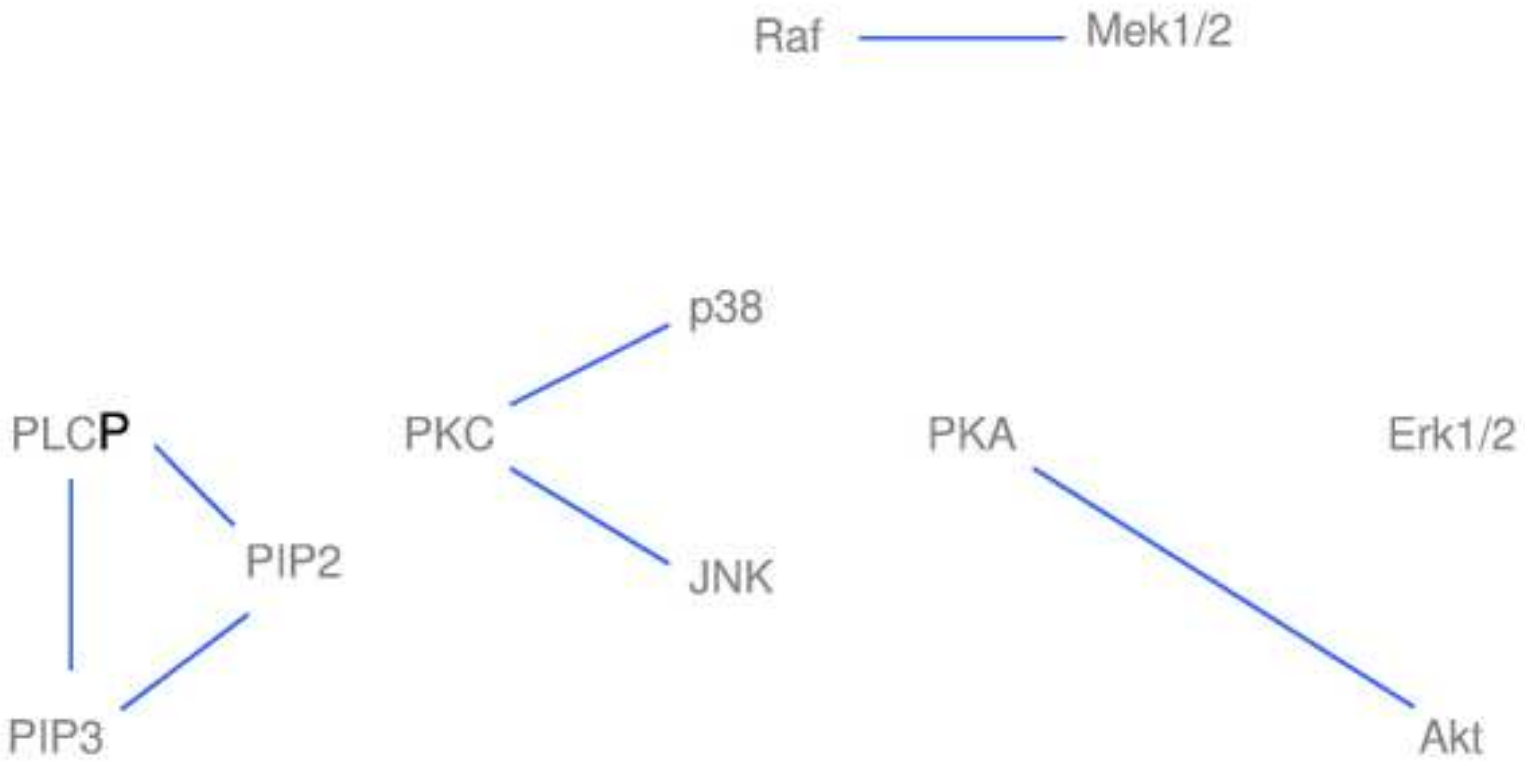
Données de cytométrie de flux : $p = 11$ protéines

$n = 902$ observations

Resultat d'une méthode d' estimation :



Graphe minimal



Test de validation de graphe : tester à partir des données que l'on n'a pas oublié d'arêtes entre 2 sommets.

Test de validation de graphe

Test de validation de graphe : "il ne manque pas d'arête dans G"

H_0 : "La structure d'indépendance conditionnelle de X est représentée par le graphe $\mathcal{G}$ ".

i.e : $a \sim b$ in $G \iff K_{ab} \neq 0$ where $K = \Sigma^{-1}$

Test de validation de graphe

Test de validation de graphe : "il ne manque pas d'arête dans G"

H_0 : "La structure d'indépendance conditionnelle de X est représentée par le graphe $\mathcal{G}$ ".

$$\text{i.e : } a \sim b \text{ in } G \iff K_{ab} \neq 0 \text{ where } K = \Sigma^{-1}$$

Test de voisinage pour chaque sommet a : "il ne manque pas de voisin au sommet a"

$$V_a = \{b : (a, b) \in \mathcal{A}\}$$

$H_{0,a}$: "$a \rightsquigarrow V_a$" contre $H_{1,a}$: " le sommet a a d'autres voisins"

Test de validation de graphe

Test de validation de graphe : "il ne manque pas d'arête dans G"

H_0 : "La structure d'indépendance conditionnelle de X est représentée par le graphe $\mathcal{G}$ ".

$$\text{i.e : } a \sim b \text{ in } G \iff K_{ab} \neq 0 \text{ where } K = \Sigma^{-1}$$

Test de voisinage pour chaque sommet a : "il ne manque pas de voisin au sommet a"

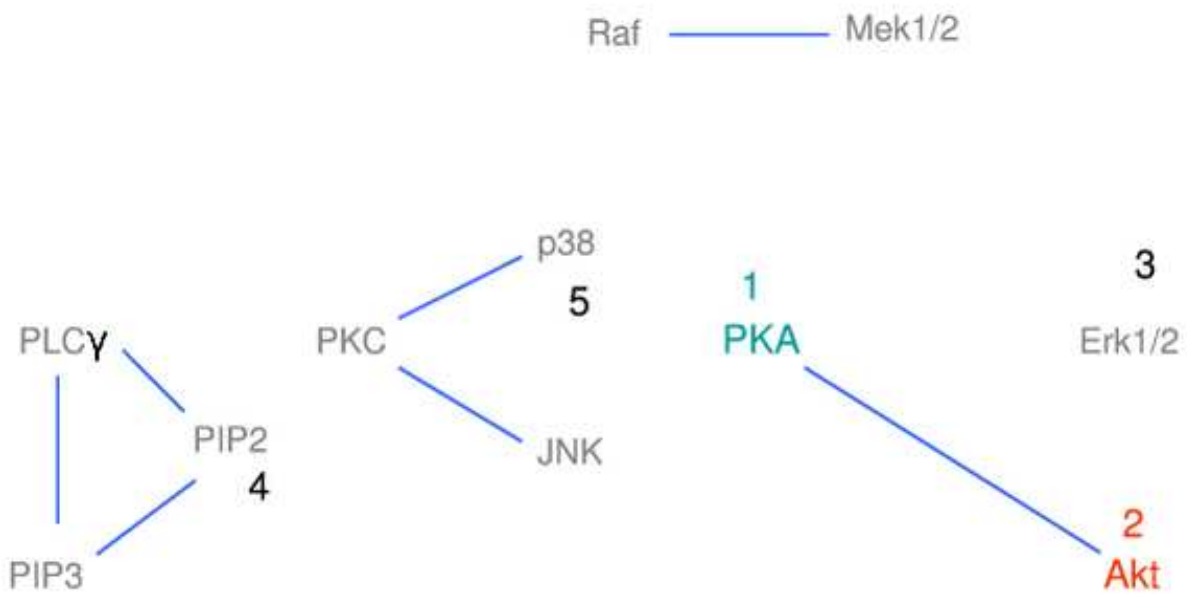
$$V_a = \{b : (a, b) \in \mathcal{A}\}$$

$H_{0,a}$: "$a \rightsquigarrow V_a$" contre $H_{1,a}$: " le sommet a a d'autres voisins"

On rejette H_0 dès que l'on rejette $H_{0,a}$ pour l'un des sommets

Test de voisinage

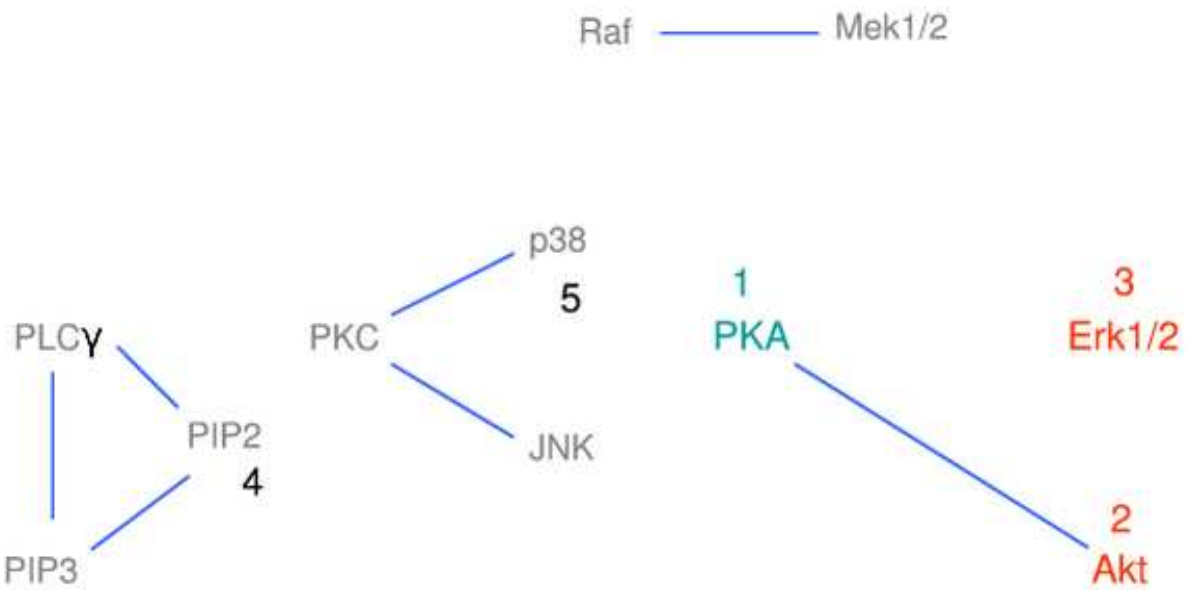
$a = 1 \quad V_a = 2$



$H_{0,a}$: "$1 \rightsquigarrow 2$" contre $H_{1,a}$: "le sommet 1 a d'autres voisins"

Test de voisinage

$a = 1 \quad V_a = 2$



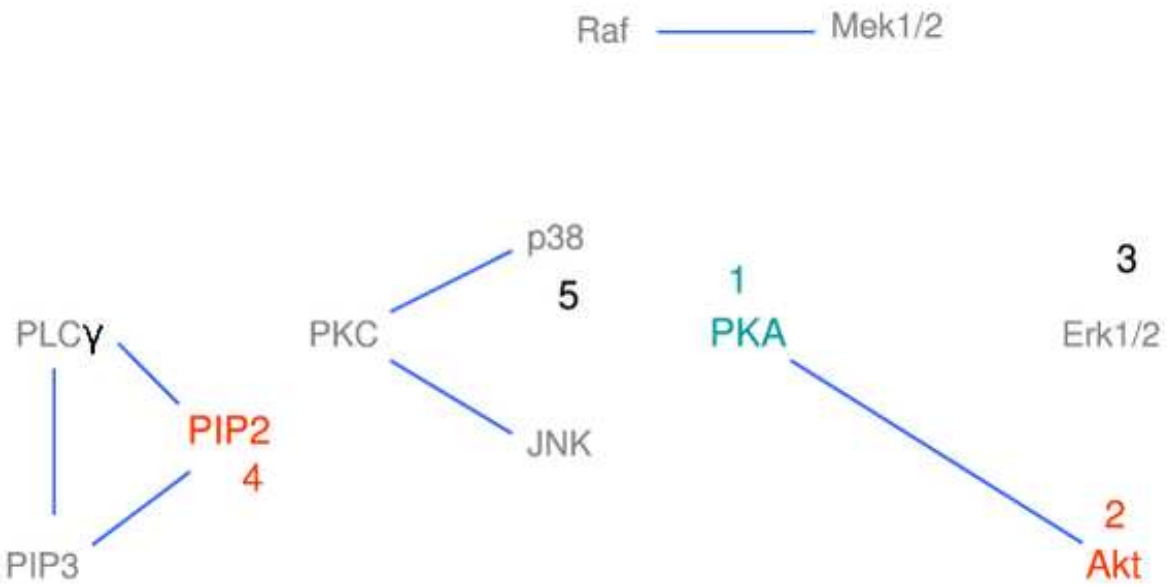
$H_{0,a}$: "$1 \rightsquigarrow 2$" contre $H_{1,a}$: " le sommet 1 a d'autres voisins"

$\Updownarrow$

$H_{0,a}$: "$1 \rightsquigarrow 2$" against $\begin{cases} H_{1,a,3} : "1 \rightsquigarrow (2,3)" \\ \end{cases}$

Test de voisinage

$a = 1 \quad V_a = 2$



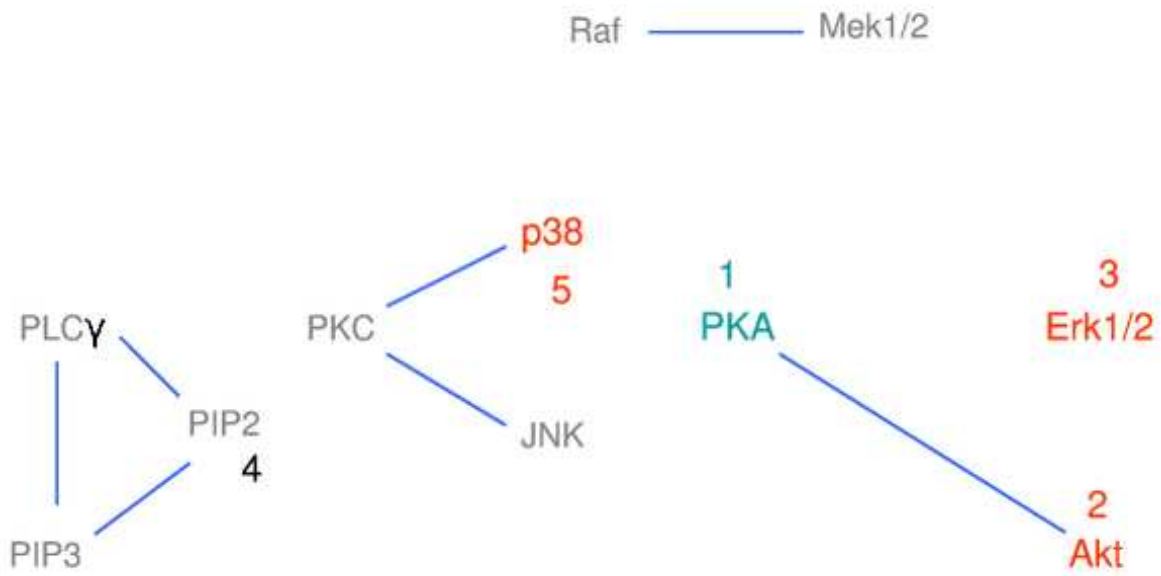
$H_{0,a}$: "$1 \rightsquigarrow 2$" contre $H_{1,a}$: " le sommet 1 a d'autres voisins"

$$\Updownarrow$$

$$H_{0,a} : "1 \rightsquigarrow 2" \text{ against } \begin{cases} H_{1,a,3} : "1 \rightsquigarrow (2,3)" \\ H_{1,a,4} : "1 \rightsquigarrow (2,4)" \end{cases}$$

Test de voisinage

$a = 1 \quad V_a = 2$



$H_{0,a}$: "$1 \rightsquigarrow 2$" contre $H_{1,a}$: "le sommet 1 a d'autres voisins"

$$\Updownarrow$$

$$H_{0,a} : "1 \rightsquigarrow 2" \text{ against } \begin{cases} H_{1,a,3} : "1 \rightsquigarrow (2,3)" \\ H_{1,a,4} : "1 \rightsquigarrow (2,4)" \\ H_{1,a,3,5} : "1 \rightsquigarrow (2,3,5)" \\ \dots \end{cases}$$

Test de voisinage

Variables Gaussiennes $\rightsquigarrow$ Régression de X_a par rapport X_{-a} :

$$X_a = \sum_{b \in \Gamma, \theta_a^a = 0} \theta_b^a X_b + \sigma_a \epsilon_a$$

Test de voisinage

Variables Gaussiennes $\rightsquigarrow$ Régression de X_a par rapport X_{-a} :

$$X_a = \sum_{b \in \Gamma, \theta_a^a = 0} \theta_b^a X_b + \sigma_a \epsilon_a$$

$$\theta_b^a = -\frac{K_{ab}}{K_{aa}} \text{ où } K = \Sigma^{-1}$$

$$a \rightsquigarrow b \iff \theta_b^a \neq 0$$

Test de voisinage pour le sommet a :

$$H_{0,a} : "a \rightsquigarrow V_a" \iff \quad "\theta^a_{\Gamma \setminus V_a} = 0"$$

Test de voisinage

Plusieurs tests : collection $\mathcal{M}_a$ de modèles m de $\Gamma \setminus \{a, V_a\}$

Pour chacun de ces sous ensembles m, tester

$$H_{0,a} : "a \rightsquigarrow V_a" \text{ contre } H_{1,a,m} : "a \rightsquigarrow \{V_a, m\}"$$

i.e

$$H_{0,a} : "\theta^a_{\Gamma \setminus V_a} = 0" \text{ contre } H_{1,a,m} : "\theta^a_{\Gamma \setminus \{V_a, m\}} = 0"$$

On rejette $H_{0,a}$ si on rejette l'un de ces tests.

- Test de $H_{0,a}$: "$\theta^a_{\Gamma \setminus V_a} = 0$" contre $H_{1,a,m}$: "$\theta^a_{\Gamma \setminus \{V_a, m\}} = 0$"

  Test de Fisher :

$$\phi_m(\mathbf{X}_a, \mathbf{X}_{-a}) = \frac{N_m \|\Pi_{V_a \cup m} \mathbf{X}_a - \Pi_{V_a} \mathbf{X}_a\|_n^2}{D_m \|\mathbf{X}_a - \Pi_{V_a \cup m} \mathbf{X}_a\|_n^2}$$

Prop

Sous $H_{0,a}$, $\phi_m(\mathbf{X}_a, \mathbf{X}_{-a}) \sim Fisher(|m|, n - |V_a| - |m|)$

- Test de $H_{0,a}$: "$\theta^a_{\Gamma\setminus V_a} = 0$" contre $H_{1,a,m}$: "$\theta^a_{\Gamma\setminus\{V_a,m\}} = 0$"

  Test de Fisher :

$$\phi_m(\mathbf{X}_a, \mathbf{X}_{-a}) = \frac{N_m \|\Pi_{V_a\cup m}\mathbf{X}_a - \Pi_{V_a}\mathbf{X}_a\|_n^2}{D_m \|\mathbf{X}_a - \Pi_{V_a\cup m}\mathbf{X}_a\|_n^2}$$

Prop

Sous $H_{0,a}$, $\phi_m(\mathbf{X}_a, \mathbf{X}_{-a}) \sim Fisher(|m|, n - |V_a| - |m|)$

- Test de voisinage pour le sommet $a \rightsquigarrow$ plusieurs tests :

$$T_\alpha = \sup_{m\in\mathcal{M}_a} \left\{\phi_m(\mathbf{X}_a, \mathbf{X}_{-a}) - \bar{F}^{-1}_{D_m,N_m}(\alpha_m)\right\}$$

On rejette $H_{0,a}$: "$\theta^a_{\Gamma\setminus V_a} = 0$" si $T_\alpha > 0$

- **Choix de la collection $\mathcal{M}_a$:**
  - Ajout des sommets 1 par 1

$$\mathcal{M}_a^1 = \{\{b\}, b \in \Gamma \setminus \{a, V_a\}\}$$

  - Modèles de plus grande dimension...

- **Choix de la collection $\mathcal{M}_a$:**
  - Ajout des sommets 1 par 1

$$\mathcal{M}_a^1 = \{\{b\}, b \in \Gamma \setminus \{a, V_a\}\}$$

  - Modèles de plus grande dimension...

- **Choix de $\{\alpha_m, m \in \mathcal{M}_a\}$ t.q niveau $=\alpha$**
  - (P_1) Bonferroni :

$$\alpha_m = \frac{\alpha}{|\mathcal{M}_a|} \Longrightarrow \mathbb{P}_{H_{0,a}}(T_\alpha > 0) \leq \alpha$$

- **Choix de la collection $\mathcal{M}_a$:**
  - Ajout des sommets 1 par 1

$$\mathcal{M}_a^1 = \{\{b\}, b \in \Gamma \setminus \{a, V_a\}\}$$

  - Modèles de plus grande dimension...

- **Choix de $\{\alpha_m, m \in \mathcal{M}_a\}$ t.q niveau $=\alpha$**
  - (P_1) Bonferroni :

$$\alpha_m = \frac{\alpha}{|\mathcal{M}_a|} \Longrightarrow \mathbb{P}_{H_{0,a}}(T_\alpha > 0) \leq \alpha$$

  - (P_2) : $\alpha_m(\mathbf{X}_{-a})$ obtenus par simulation tel que :

$$\mathbb{P}_{H_{0,a}}\left(\sup_{m \in \mathcal{M}_a}\left\{\phi_m(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a}) - \bar{F}^{-1}_{D_m, N_m}(\alpha_m(\mathbf{X}_{-a}))\right\} > 0 \mid \mathbf{X}_{-a}\right) = \alpha$$

- **Choix de la collection $\mathcal{M}_a$:**
  - Ajout des sommets 1 par 1

$$\mathcal{M}_a^1 = \{\{b\}, b \in \Gamma \setminus \{a, V_a\}\}$$

  - Modèles de plus grande dimension...

- **Choix de $\{\alpha_m, m \in \mathcal{M}_a\}$ t.q niveau $=\alpha$**
  - (P_1) Bonferroni :

$$\alpha_m = \frac{\alpha}{|\mathcal{M}_a|} \Longrightarrow \mathbb{P}_{H_{0,a}}(T_\alpha > 0) \leq \alpha$$

  - (P_2) : $\alpha_m(\mathbf{X}_{-a})$ obtenus par simulation tel que :

$$\mathbb{P}_{H_{0,a}}\left(\sup_{m \in \mathcal{M}_a}\left\{\phi_m(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a}) - \bar{F}^{-1}_{D_m,N_m}(\alpha_m(\mathbf{X}_{-a}))\right\} > 0 \mid \mathbf{X}_{-a}\right) = \alpha$$

En fait :
$\alpha_m(\mathbf{X}_{-a})$= le α-quantile, conditionnellement à $\mathbf{X}_{-a}$ de

$$\inf_{m \in \mathcal{M}_a} \bar{F}_{D_m,N_m}\left(\phi_m(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a})\right)$$

Puissance du test

Theorème

T_α est le test avec la collection $\mathcal{M}_a^1$ et la procédure P_1. Soient (n, p, d_a) tels que :

$$n - d_a - 1 \geq \left[10 \log \left(\frac{p - d_a - 1}{\alpha} \right) \vee 21 \log (1/\delta) \right]$$

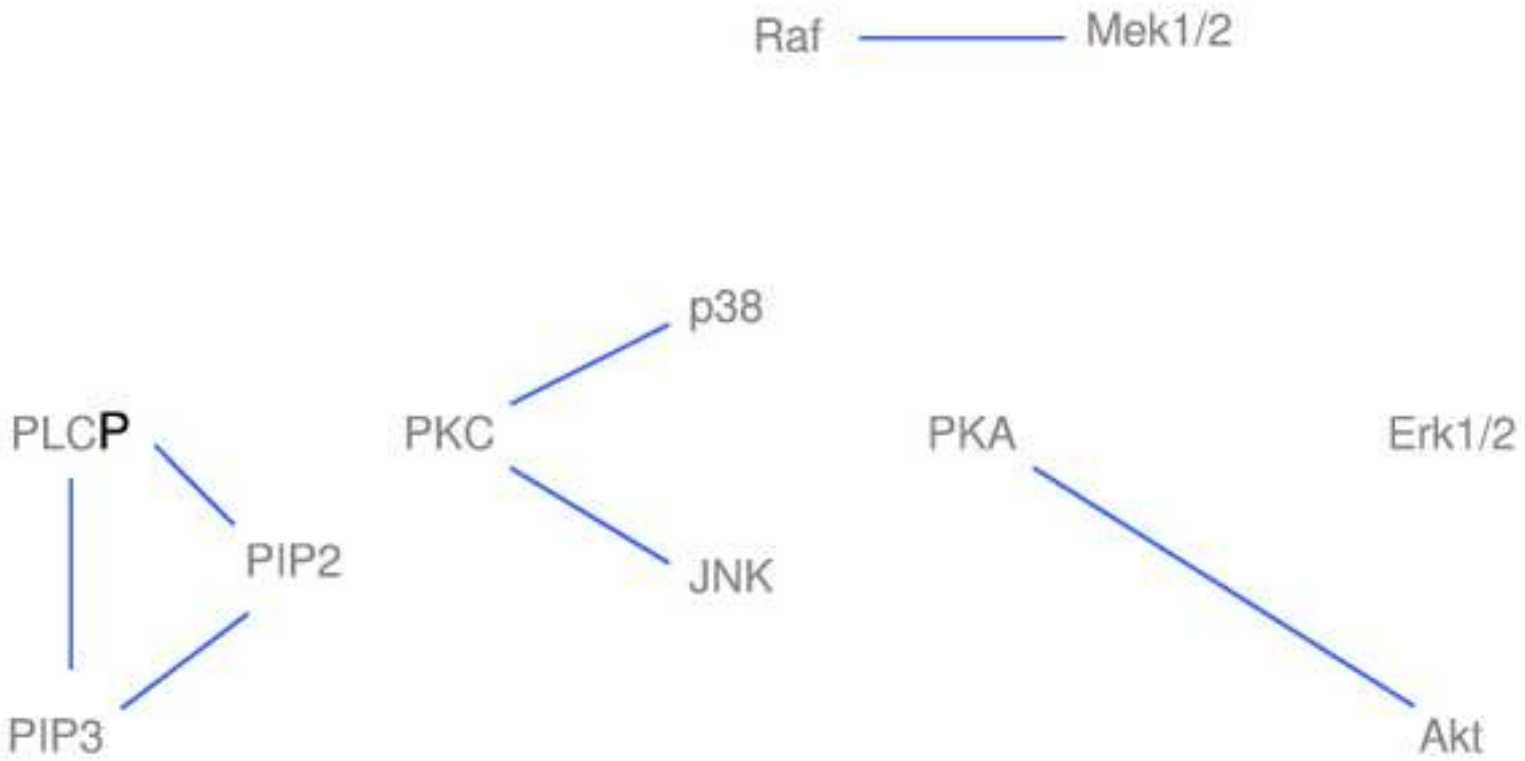
Soit

$$\rho^2_{n-d_a, p-d_a} = \frac{C_1}{n - d_a} \log \left(\frac{p - d_a - 1}{\alpha \delta} \right)$$

Alors $\mathbb{P}_{\theta^a} (T_\alpha > 0) \geq 1 - \delta$ pour tout θ^a tel que

$$\exists b \in \Gamma \setminus \{a, V_a\} : \frac{\text{var}(X_a | X_{V_a}) - \text{var}(X_a | X_{V_a \cup \{b\}})}{\text{var}(X_a | X_{V_a \cup \{b\}})} \geq \rho^2_{n-d_a, p-d_a}$$

Test de validation du graphe minimal $\mathcal{G}_1$ à partir des données de Sachs ($n = 902$ observations sur les $p = 11$ protèines):

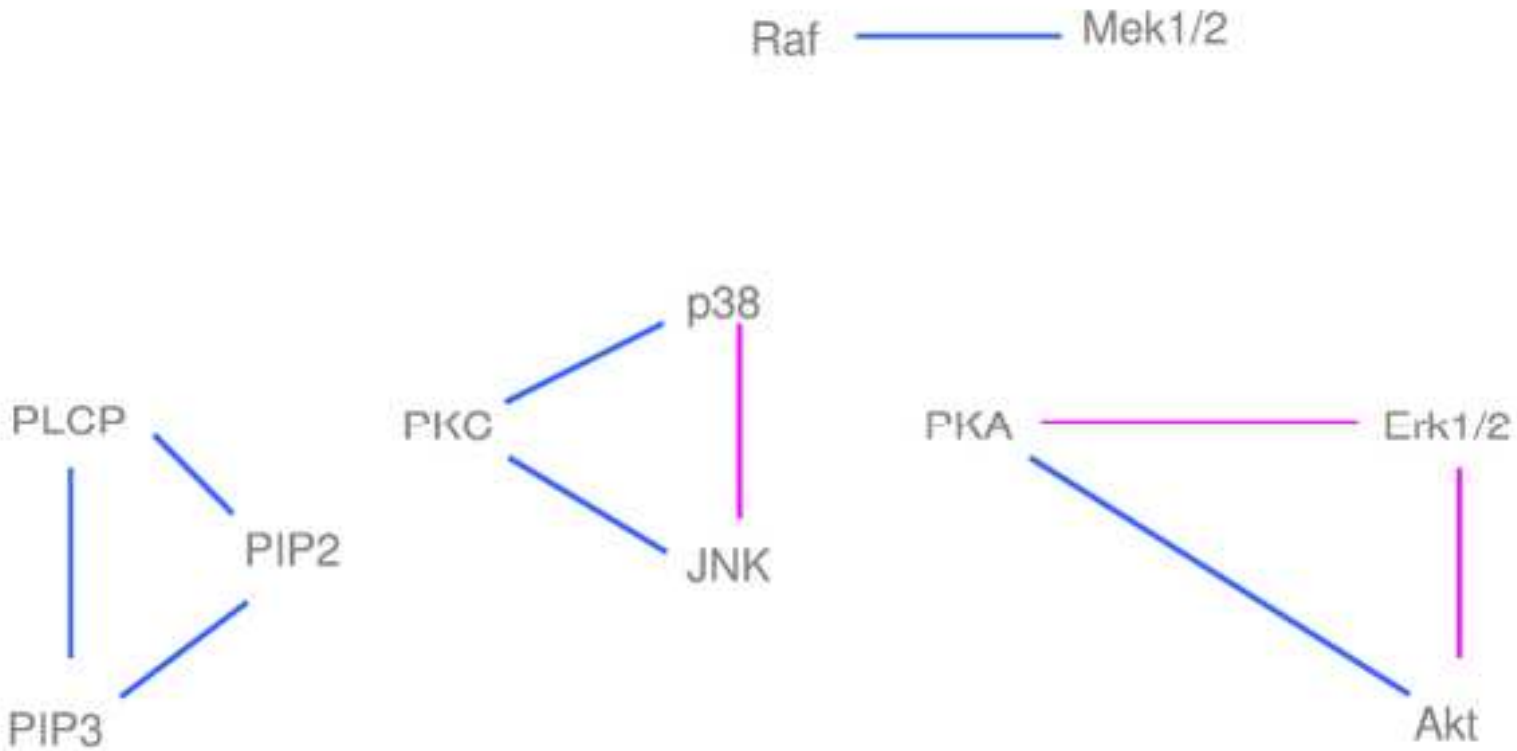


Graphe $\mathcal{G}_1$

H_0 : " La structure d'indépendance conditionnelle de X est représentée par le graphe $\mathcal{G}_1$"

On rejette H_0 au niveau 5%

Test de validation du graphe minimal $\mathcal{G}_2$ à partir des données de Sachs



Graphe $\mathcal{G}_2$

H_0 : " La structure d'indépendance conditionnelle de X est représentée par le graphe $\mathcal{G}_2$"

On accepte H_0

Influence de n

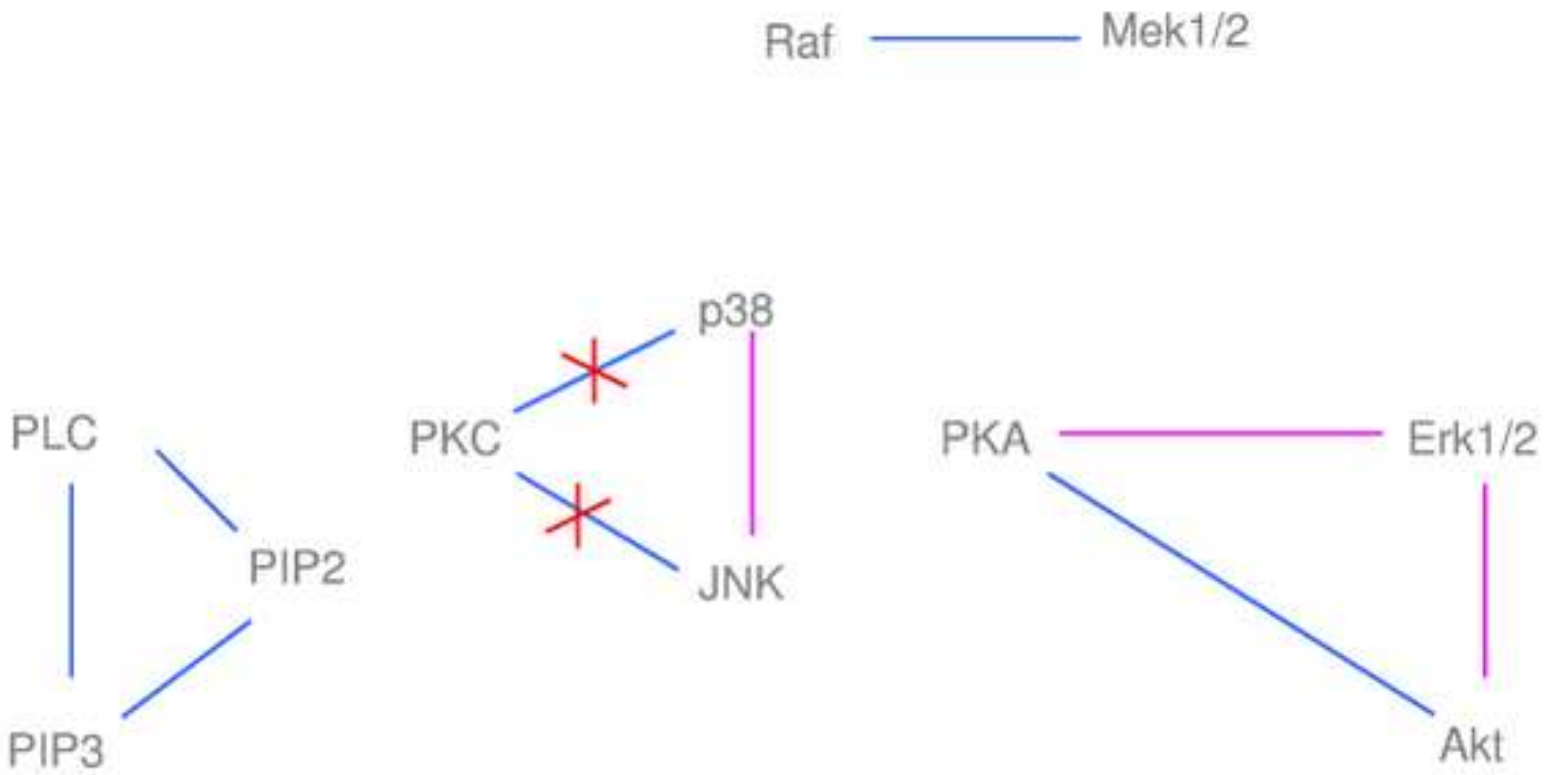
$n = 10, 15, 20$

1000 simulations : Pour chaque simulation s , on simule n observations d'un vecteur gaussien X^s telles que la structure d'indépendance conditionnelle de X^s soit représentée par le graphe $\mathcal{G}_2$

On estime le niveau en testant :

H_0 : La structure d'indépendance conditionnelle de X^s est représentée par le graphe $\mathcal{G}_2$

n	niveau
10	0.032
15	0.036
20	0.033



Graphe $\mathcal{G}_2^-$: graphe $\mathcal{G}_2$ auquel on a retiré 2 arêtes.

On estime la puissance en testant :

H_0 : La structure d'indépendance conditionnelle de X^s est représentée par le graphe $\mathcal{G}_2^-$

n	puissance
10	0.49
15	0.86
20	0.97

Messages clés

- construction d'une procedure de "test de validation de graphe"
- proprietés : niveau controlé; test puissant
- peux suggérer de nouvelles interactions entre les gènes

Merci de votre attention...