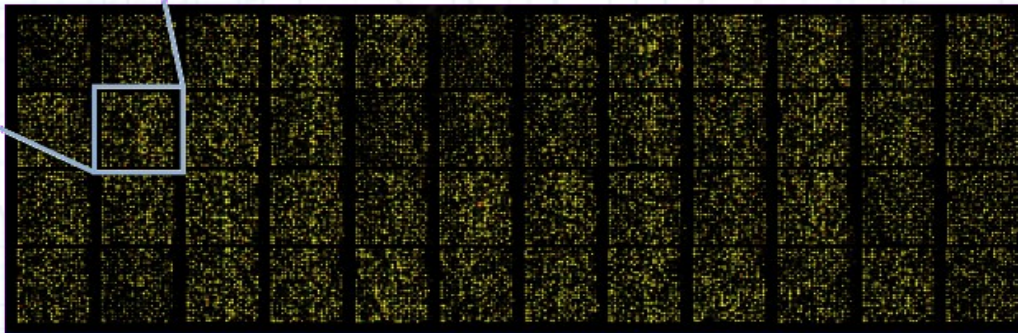
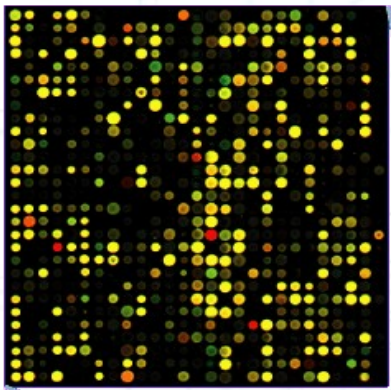


Utilisation des couvertures de Markov pour la sélection de variables ; application à l'obésité

David Dernoncourt

INSERM UMRS 872 Equipe NUTRIOMIQUE
Centre de Recherche des Cordeliers
Encadrants: JD Zucker, K Clément, B Hanczar

Contexte : analyse de données d'expression

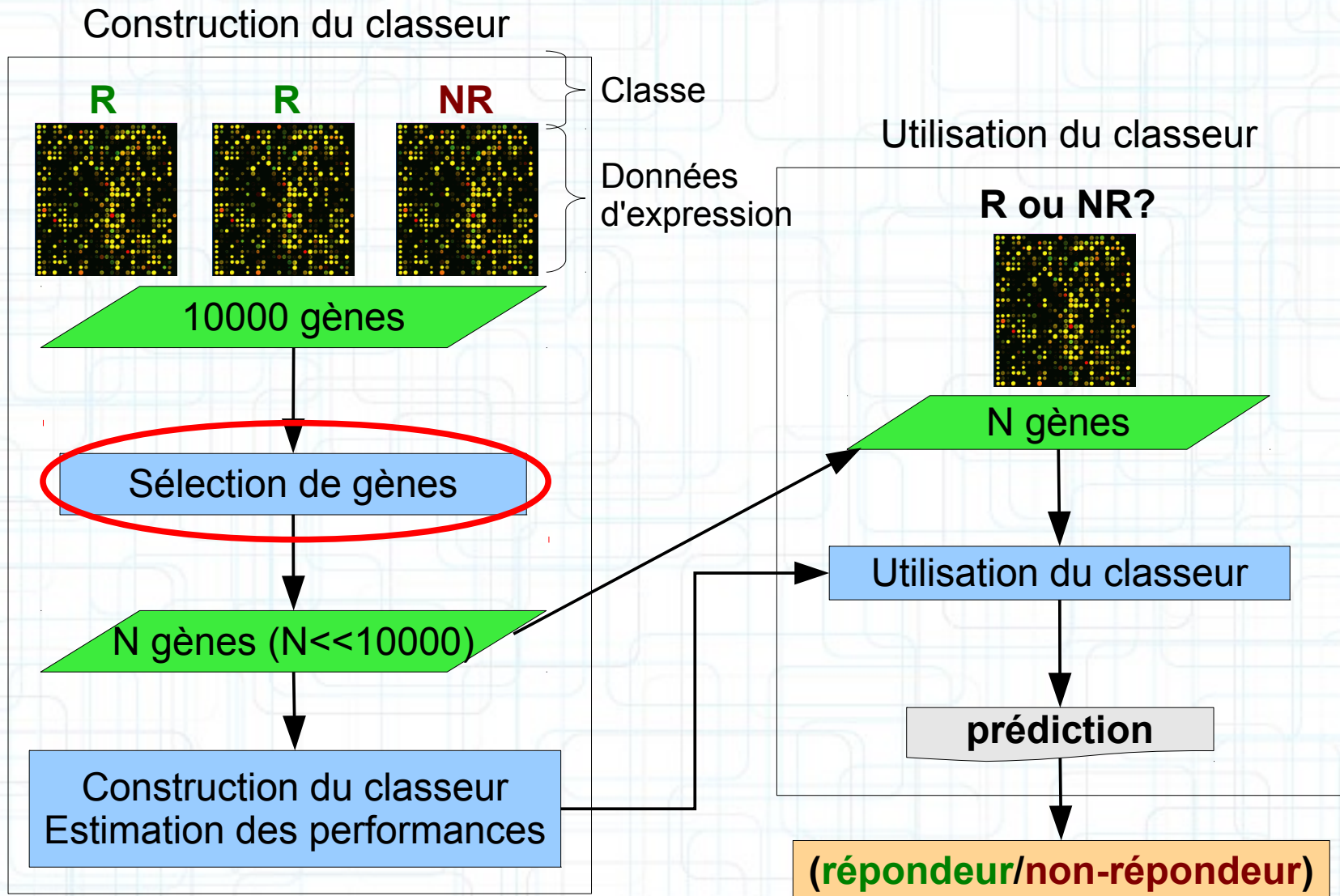


- Données d'expression génique (puces à ADN):
 - haute dimension (~10000 gènes)
 - coûteux => peu d'observations (~100 patients)
- Une utilisation: **prédiction** d'un phénotype ("classe"):
 - type de cancer¹, survie²
 - **répondeur/non-répondeur**
 - par exemple: **reprise de poids après régime**

1: Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* 2002; 18: 39–50.

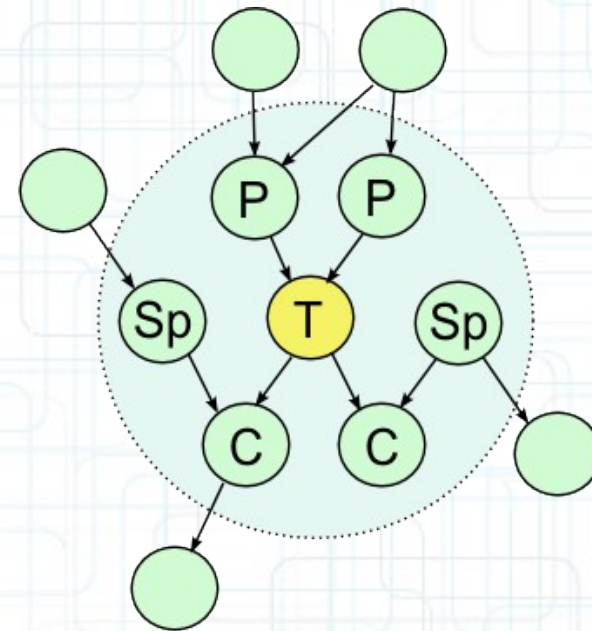
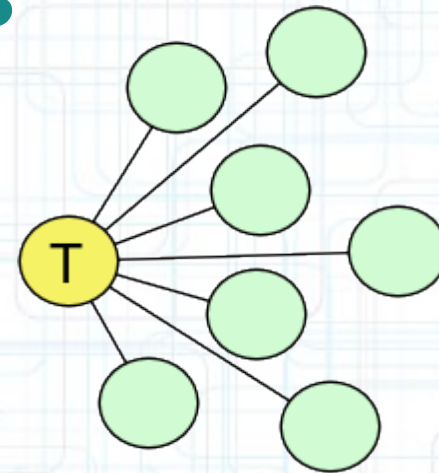
2: Van de Vijver MJ, He YD, van't Veer LJ et al. A gene-expression signature as a predictor of survival in breast cancer. *New Eng J Med* 2002; 347: 1999–2009

Contexte : la prédiction à partir de données d'expression



Problématique : la sélection de gènes

- Nombreuses méthodes de **sélection de gènes** utilisées¹ :
 - univariés : t-test...
 - multivariés : t ajusté sur corrélation²...
- Koller & Sahami, 1996;
Tsamardinos & Aliferis, 2003³:
“*the **theoretically** optimal solution to the feature selection problem is the Markov blanket of the target variable (T)*”
- Mais... aucune étude sur données puces avec moins de 200 obs (>99% des cas⁴)



1: Saeys Y., Inza I., Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007; 23: 2507–17.

2: Zuber V., Strimmer, K. Gene ranking and biomarker discovery under correlation. *Bioinformatics* 2009; 25: 2700–7.

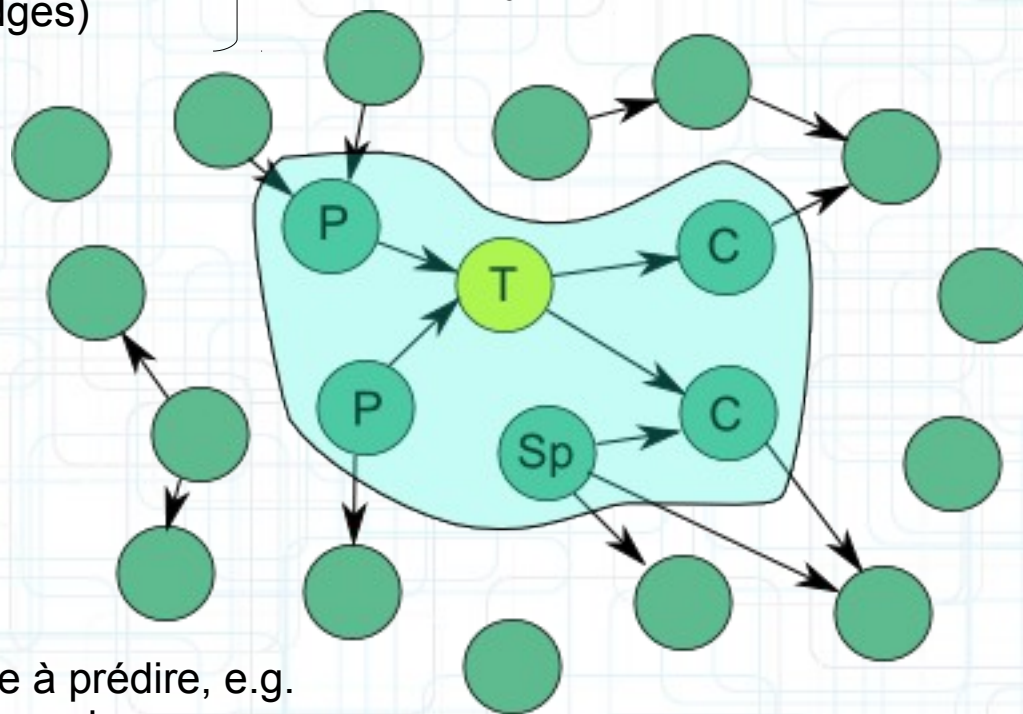
3: Tsamardinos I., Aliferis CF. Towards principled feature selection: relevancy, filters and wrappers. In: *AISTATS* 2003.

4: *Gene Expression Atlas* et *NCBI Gene Expression Omnibus*

Réseau bayésien et couverture de Markov

noeuds/sommets (variables)
+ arc (directed edges)

Réseau bayésien



T (target): variable à prédire, e.g.
répondeur/non-répondeur

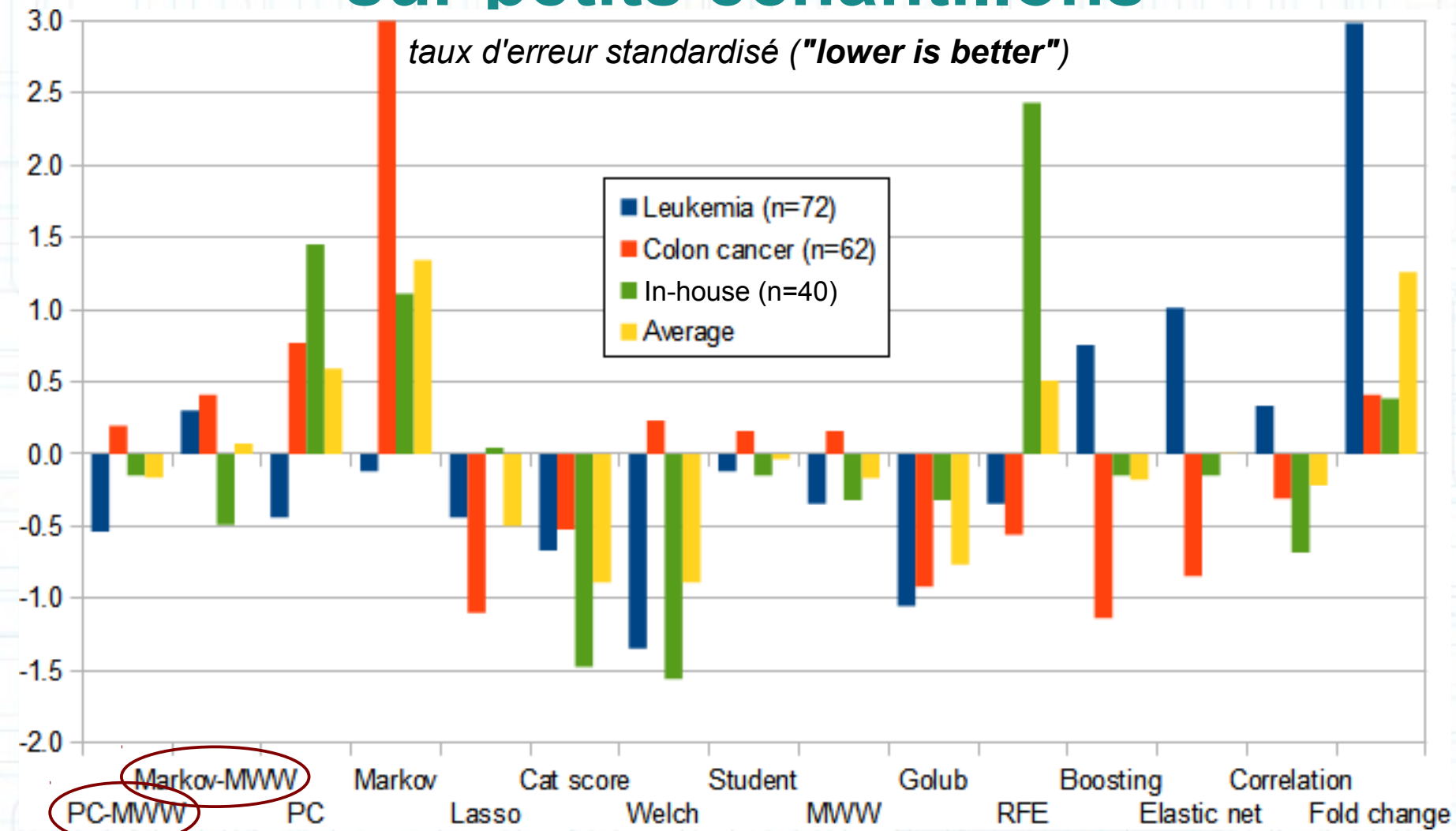
C (child): enfant
P: parent
Sp (spouse): coparent

Couverture de Markov de T

Résultats de master

- Problème: calcul de MB de manière exacte et efficiente, précision de la prédiction en résultant
- Point de départ: algorithme de filtre par MB correspondant à l'état de l'art: IPC-MB
- Test (données artificielles, données puces publiques, données du laboratoire nutriomique) vs autres filtres
- Modification de l'algorithme:
 - "uncached variable removal" (correction mineure)
 - ↓ nécessité de discrétisation (moins de perte d'information)
- Principaux résultats:
 - bonne prédiction sur réseaux bayésiens artificiels simples
 - bonne prédiction sur données puces >200 observations

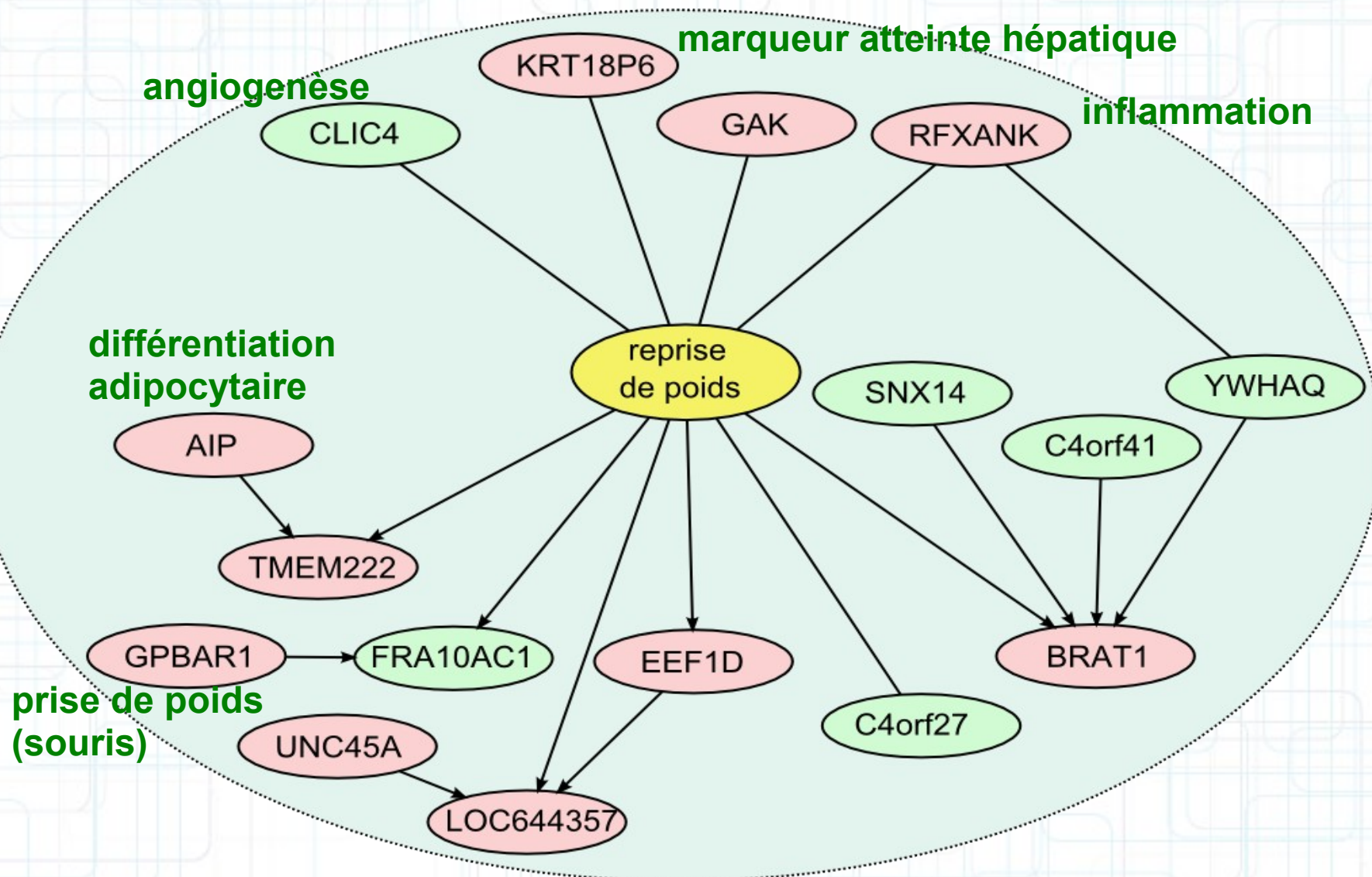
Resultats : taux d'erreur sur petits échantillons



- Algorithme modifié: amélioration de la prédiction sur petits échantillons, de mauvais à acceptable

Résultats : MB de prédiction de reprise du poids

- Couverture de Markov : 16 gènes sur 13078 ; n=40



Perspectives: filtre MB

- Améliorer encore l'algorithme:
 - généralisation:
 - classification avec > 2 classes
 - données qualitatives
 - score de pertinence des variables sélectionnées
(MB est \pm le seul filtre qui ne fournit pas un tel score)
- Application à d'autres données:
 - données d'expression (transcriptomiques, métagénomiques)
 - ou autres, notamment des données clinico-biologiques

Perspectives: stabilité/robustesse de la sélection

- Le problème de la stabilité de la sélection de gènes:
 - un objectif de la sélection et classification est de révéler des gènes candidats.
 - d'une étude à l'autre, traitant de la même tâche de prédiction, les recouvrements entre les gènes du modèle final est en règle proche de zéro¹.
 - il est probable qu'une sélection de gènes plus stable sortirait de meilleurs gènes candidats.
- Pour aborder ce problème:
 - évaluation de la stabilité des méthodes de sélection existantes (dont la couverture de Markov)
 - essayer de construire une méthode plus stable

1: Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, Eytan Domany. Outcome signature genes in breast cancer: is there a unique set?. *Bioinformatics* 2005; 21: 171–8.

Remerciements

INSERM U872 Nutriomique

www.crcjussieu.fr

Jean-Daniel Zucker

Blaise Hanczar

Karine Clément

Aurélie Cotillard

Edi Prifti

Meriem Abdennour

Henri Hooton

Flavien Jacques

Véronique Pelloux

Ramzi Temanni

Salwa Rizkalla

Joan Tordjman

Nadine le Pontois

&

**Les autres membres de l'équipe (clinique,
recherche)**

UPMC
SORBONNE UNIVERSITÉS

Inserm

Institut national
de la santé et de la recherche médicale

ASSISTANCE
PUBLIQUE  HÔPITAUX
DE PARIS