Biostatistique pour les données "omiques"

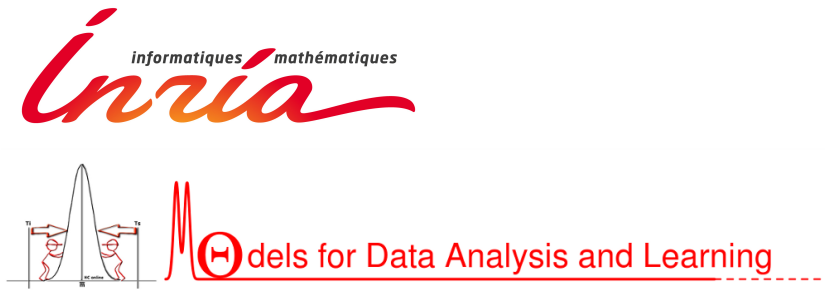
Guillemette Marot
Univ. Lille 2, EA 2694 & Inria MODAL

6 janvier 2017

Plan

Compétences équipes

Santé publique : épidémiologie et qualité des soins
Membres : cliniciens, épidémiologistes et méthodologistes
Domaines : biostatistique, informatique médicale et économie de la santé.



Equipe-Projet : modèles pour l'analyse de données et l'apprentissage
Membres : statisticiens de Lille 1, Lille 2 et Lille 3
Spécialité : modèles génératifs

Univ. Lille Droit et Santé, EA 2694

Equipe d'accueil 2694

Santé publique : épidémiologie et qualité des soins

80 Membres dans l'équipe

- 47 enseignants chercheurs / Chercheurs
- 3 praticiens hospitaliers (PH)
- 18 doctorants
- 11 ingénieurs à temps partiel
- 1 secrétaire à mi-temps

Univ. Lille Droit et Santé, EA 2694

Axe I : Epidémiologie

- Obésité : Evaluation des programmes de Santé communautaire, Activité physique et obésité des enfants
- Maladies Transmissibles : modèles de transmission, coût-efficacité, vacccinations
- En développement : prévention et complication des infections de la mère et de l'enfant

Axe II : Qualité des soins

- Innovations technologiques (évaluation d'outils d'aide à la prescription, évaluation des dispositifs médicaux)
- Outils d'aide à la décision (détection des effets indésirables des médicaments, scores, base de données pédiatrie)
- Evaluation des pratiques Professionnelles (Registre national des arrêts cardiaques RéAc)
- En développement : méthodologie d'évaluation des technologies de santé

CERIM : centre d'études et de recherche en informatique médicale

Lieu au pôle recherche de la faculté de médecine qui rassemble principalement des membres de l'EA2694 travaillant en informatique médicale et e-santé ou biostatistique.

Thèmes de recherche des "méthodologistes" du CERIM :

- traitement des données manquantes
- statistiques de scan
- analyse des données de survie
- modèles non linéaires mixtes
- analyse des données génomiques

Univ. Lille Droit et Santé, EA 2694

Implication forte de certains membres du CERIM pour encadrer des ingénieurs statisticiens de deux plateformes :

- plateforme méthodologique du CHRU localisée à la Maison Régionale de la Recherche Clinique
- plateforme de bioinformatique et bioanalyse bilille
  (8 tutelles : Univ. Lille 1, Univ. Lille 2, CNRS, Inserm, Inria, CHRU, Institut Pasteur de Lille, Institut de Recherche contre le Cancer de Lille)

MODAL

Models for Data Analysis and Learning

- 7 membres permanents (5 Univ. Lille 1 Painlevé, 2 Univ. Lille 2 EA 2694)
- 3 membres associés (Univ. Lille 3, Univ. Lyon 2, PGXIS)
- 5 doctorants
- 3 à 5 ingénieurs (en comptant les ingénieurs InriaTech ou EA 2694)
- 1 assistante d'équipe

Objectif 1 : Conception de modèles dans un espace de départ

Objectif 2 : Conception de modèles dans un espace à noyaux

Objectif 3 : Visualisation à travers les modèles

Objectif 4 : Applications biologiques et modèles

Inria, MODAL

Thèmes de recherche des membres permanents de MODAL :

- modèles de mélange
- données fonctionnelles
- approches PAC-bayésiennes
- segmentation à noyaux
- sélection de modèles
- sélection de variables
- visualisation

Plan

1. Présentation des équipes
   - Univ. Lille Droit et Santé, EA 2694
   - Inria, MODAL

2. Biostatistique pour les données omiques
   - Introduction
   - Analyse différentielle en transcriptomique
   - Sélection de marqueurs génomiques
   - Intégration de données -omiques
   - Conclusion - perspectives

Introduction

Des données volumineuses !

Exemple emblématique : **projet génome humain** 1990 $\rightarrow$ 2003
13 ans pour séquencer le génome de référence avec 2,7 Mds dollars et une vingtaine de labos.
Aujourd'hui, séquençage d'un génome complet dans un seul labo en 3 jours.

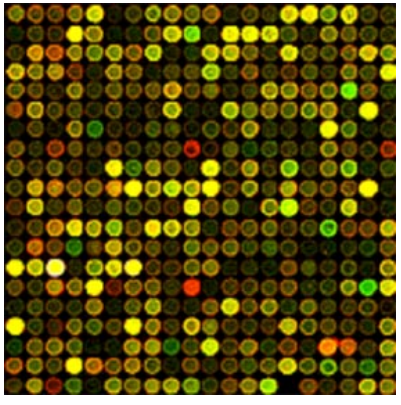
Plusieurs niveaux à étudier

- génomique : modifications de l'ADN (ex : SNP, CNV)
- transcriptomique : différence d'expression = différence de la quantité d'ARNm
- épigénomique : mécanismes moléculaires concernant le génome ainsi que l'expression des gènes, qui peuvent être influencés par l'environnement et l'histoire individuelle. On regarde par exemple l'état de la chromatine, les modifications d'histones.

Introduction

Des technologies qui permettent d'étudier plusieurs milliers voire millions de marqueurs simultanément

Puces à ADN

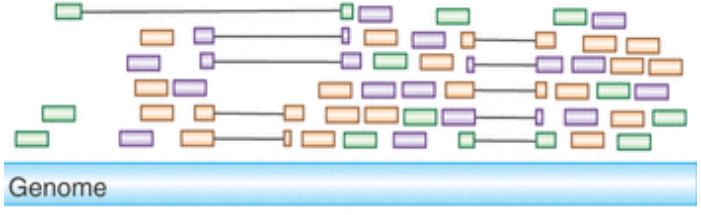


source :

bio.cse.ohio-state.edu/MicroarrayDesigner/

Mesures d'intensités lumineuses

Séquençage à haut débit



source :Haas and Zody, Nature Biotechnology (2010)

Comptages de lectures

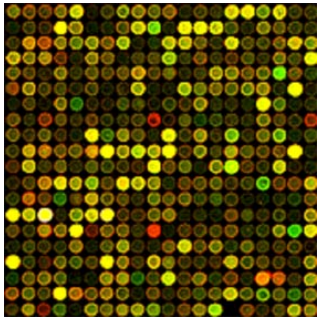
Normalisations et modélisations spécifiques à chaque technologie

Plan

Analyse différentielle en transcriptomique

Transcriptomique

Mesure de l'expression de plusieurs milliers de gènes simultanément.



But : trouver des gènes différentiellement exprimés

Problème statistique de ces expériences : grande dimension ($n \ll T$)

- Beaucoup de tests (plusieurs milliers : T gènes)
- Très peu d'individus (quelques dizaines : n)

Transcriptomique

Modèle statistique

$$y_{gcr} = m_{gc} + e_{gcr}$$

y_{gcr} : niveau d'expression pour le gène g, le réplicat r et la condition c .
m_{gc} : effet du gène g dans la condition c (effet fixe).

$$e_{gcr} \sim \mathcal{N}(0, \sigma^2_{gc})$$

Variances σ^2_{gc} peuvent varier à la fois par gène g et par condition c.

La modélisation de la variance est une étape importante dans la construction de la statistique de test.

Transcriptomique

Modélisation des variances σ^2_{gc}

Une variance estimée pour chaque gène :

- Manque de sensibilité dû au manque d'information.
- Trop peu de gènes détectés.

Variance commune pour tous les gènes :

- Beaucoup de gènes détectés.
- Beaucoup de faux positifs.

Approches de rétrécissement :

Permettent de gagner de la sensibilité tout en gardant le nombre de faux positifs assez bas.

Transcriptomique

Rétrécissement

$$\tilde{\theta_g} = \widehat{\theta_c} + b(\widehat{\theta_g} - \widehat{\theta_c})$$

avec $\tilde{\theta_g}$ = estimateur "rétréci"
$\widehat{\theta_c}$ = estimateur de la population moyen
$\widehat{\theta_g}$ = estimateur empirique gène à gène
b = facteur de rétrécissement

$b = 1 \Rightarrow \tilde{\theta_g} = \widehat{\theta_g}$
$b = 0 \Rightarrow \tilde{\theta_g} = \widehat{\theta_c}$

Modélisation de la variance

Modèle structural mixte sur les variances : $\theta = \ln(Var(y_{gcr}))$
(Jaffrézic et al., 2007 ; Foulley et al., 1992)

$$\ln(\sigma^2_{gc}) = \mu_c + \delta_{gc}$$

$$\delta_{gc} \sim \mathcal{N}(0, \tau^2_c)$$

μ_c : condition (effet fixe)
δ_{gc} : effet du gene g dans la condition c (effet aléatoire)

$\Rightarrow$ Flexibilité : une variance pour chaque gène dans chaque condition.
Statistique de Welch.

Package SMVar

Transcriptomique

Approches bayésiennes empiriques meilleures que les approches précédentes pour tester l'expression différentielle des gènes.

Limma est certainement le package le plus utilisé par les personnes qui analysent les puces d'expression.

SMVar, package dans lequel est implémenté le modèle structural pour les variances pour les études de gènes différentiellement exprimés, est particulièrement utile pour des variances hétérogènes entre conditions.

Plan

Sélection de marqueurs génomiques

Sélection de marqueurs génomiques

Classification de profils avec sélection de positions

Projet MPAGenomics

Collaboration avec la plateforme de génomique fonctionnelle et structurale de Lille 2 (M. Figeac) et le laboratoire d'hématologie de Lille (C. Preudhomme)

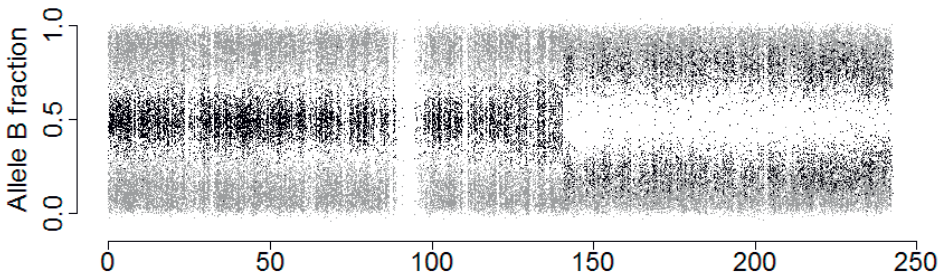
Puces Affymetrix SNP6.0

Génération de puces permettant d'étudier à la fois les CNV et les pertes d'hétérozygotie (**LOH** "AB" $\Rightarrow$ "AA" or "AB")

Sélection de marqueurs génomiques

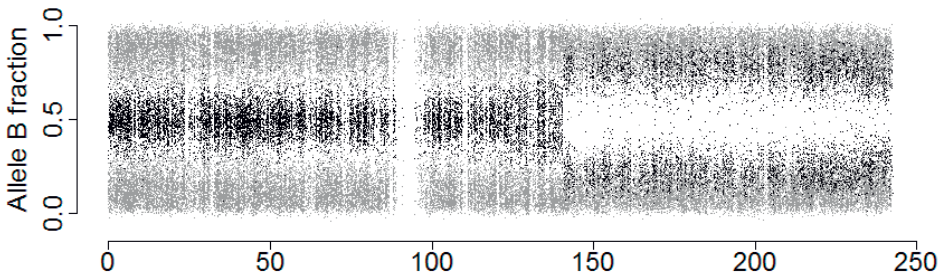
Avant le projet, génotypage classique = classification supervisée de signaux avec un échantillon d'apprentissage constitué d'individus du projet HapMap

$\Rightarrow$ transformation des signaux normalisés en statuts "AA","AB","BB" (fractions d'allele B 0, 1/2, 1)



Sélection de marqueurs génomiques

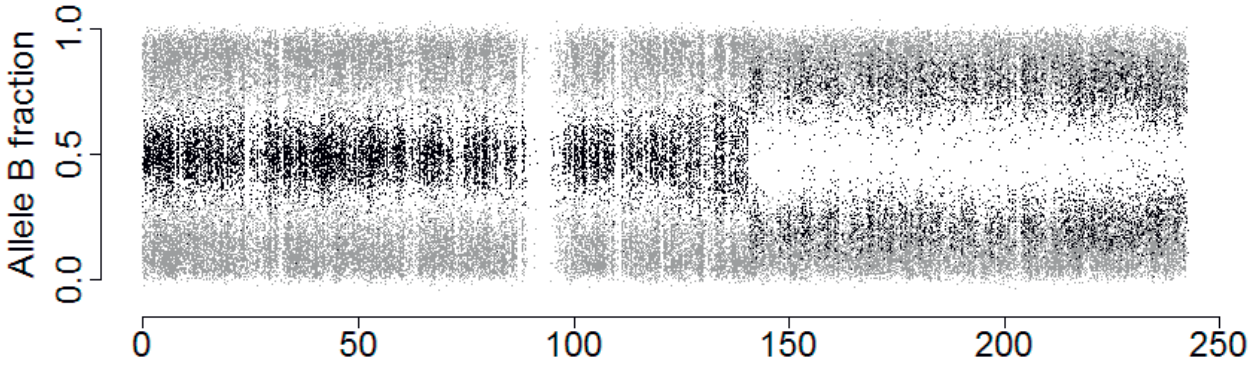
Avant le projet, génotypage classique = classification supervisée de signaux avec un échantillon d'apprentissage constitué d'individus du projet HapMap
$\Rightarrow$ transformation des signaux normalisés en statuts "AA","AB","BB" (fractions d'allele B 0, 1/2, 1)



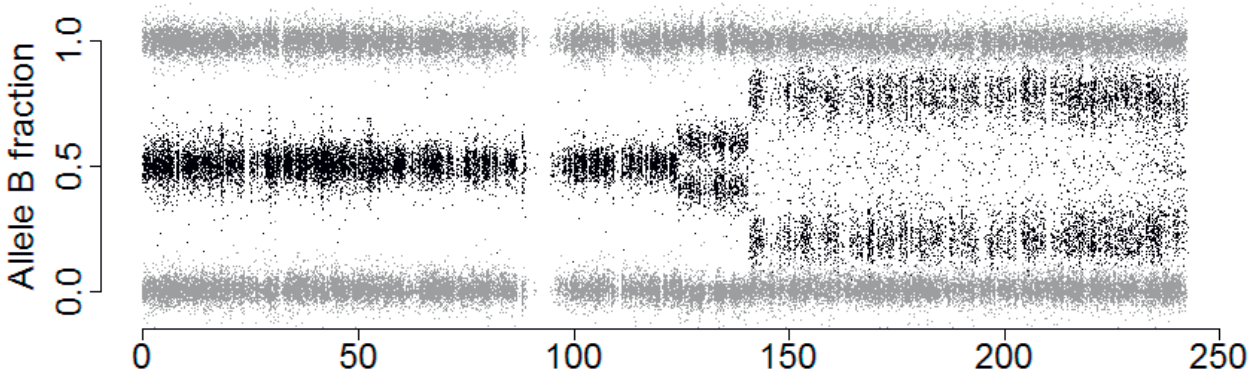
Problèmes avec les échantillons tumoraux :

- Nombre de copies =3 $\Rightarrow$ la méthode de génotypage ne fournit pas les génotypes "AAA", "BBB", "AAB" or "ABB".
- On pourrait aussi observer un mélange de signaux types provenant à la fois de cellules normales et tumorales.

Sélection de marqueurs génomiques



Normalisation TumorBoost (Bengtsson, 2010)



Sélection de marqueurs génomiques

MPAGenomics

Intégration de la normalisation Tumorboost avec des méthodes de classification en grande dimension avec sélection de variables

Problème de la grande dimension

Problème de multicolinéarité

Nécessité de faire soit une sélection soit un regroupement de variables pour une meilleure interprétation, il ne suffit pas de pouvoir classer un nouveau patient

$\Rightarrow$ notion de **sparsité** (très peu de coefficients différents de 0) ou de **classification croisée**

Sélection de marqueurs génomiques

Problèmes sparses bien connus en régression.

Régression via Lasso

Minimisation de $\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{T} |\beta_j|$

- Equivalent à minimiser la somme des carrés avec une contrainte $\sum_{j=1}^{T} |\beta_j| \leq s$
- Pénalités L1 ont des avantages à la fois statistiques et computationnels.

Sélection de marqueurs génomiques

Premiers essais sur des profils soit de CN soit de fraction d'allèle B symétrisés :

Deux hypothèses envisagées :

1) Les patients d'un même sous-groupe partagent des **marqueurs communs** (CNV, SNP) qui permettent de comprendre un mécanisme de dormance tumorale ou de prédire la rechute d'un patient :

2) Les **profils** sont très **hétérogènes**, même à l'intérieur d'un sous-groupe.

Sélection de marqueurs génomiques

1) Cas de marqueurs communs :
Les régressions pénalisées atteignent leur limite en ultra-grande dimension $\Rightarrow$ thèse de Quentin Grimonprez (2016) : Multi-Layer Group Lasso

2) Cas de profils très hétérogènes :
Il est nécessaire d'analyser automatiquement et individuellement chaque profil pour trouver les anomalies, certaines pouvant être rares dans la population étudiée $\Rightarrow$ travaux pour mieux calibrer les paramètres (Grimonprez et al., 2014) ou pour améliorer la segmentation (Celisse et al., 2016)

Multi-layer group lasso (Grimonprez et al., 2016)

Contexte :

- sélection de variables en grande dimension à l'aide de procédures de régression régularisée
- présence de redondance entre variables explicatives (ex : plusieurs sondes d'un même gène).
- réponse univariée quantitative (ex : mesure du nombre de globules blancs)

Hypothèse : Parmi les variables candidates, seul un petit nombre est réellement pertinent pour expliquer la réponse.

Multi-layer group lasso (Grimonprez et al., 2016)

Contexte :

- sélection de variables en grande dimension à l'aide de procédures de régression régularisée
- présence de redondance entre variables explicatives (ex : plusieurs sondes d'un même gène).
- réponse univariée quantitative (ex : mesure du nombre de globules blancs)

Hypothèse : Parmi les variables candidates, seul un petit nombre est réellement pertinent pour expliquer la réponse.

Problèmes :

- les approches classiques de type Lasso voient leurs performances se dégrader lorsque la redondance croît, puisqu'elles ne tiennent pas compte de cette dernière.
- regrouper au préalable ces variables peut pallier ce défaut, mais nécessite usuellement la calibration de paramètres supplémentaires ou la connaissance des groupes.

Multi-layer group lasso (Grimonprez et al., 2016)

Approche proposée :

- basée sur Classification Ascendante Hiérarchique (CAH) + Group Lasso
- groupes candidats issus potentiellement de différents niveaux de la CAH à paramètre de régularisation fixé
- choix des groupes par procédure de tests multiples

Originalité :

Exploitation de la structure hiérarchique de la CAH et des pondérations dans le Group-lasso pour réduire la complexité algorithmique induite par la flexibilité liée à la possibilité de choisir des groupes issus de différents niveaux de la CAH.

Plan

Intégration de données -omiques

Intégration de données

Points de vue complémentaires :

- Classification (non supervisée, supervisée)

Intégration de données

Points de vue complémentaires :

- Classification (non supervisée, supervisée)
- Sélection de variables, identifications de biomarqueurs
- Prédiction de phénotypes

Intégration de données

Points de vue complémentaires :

- Classification (non supervisée, supervisée)
- Sélection de variables, identifications de biomarqueurs
- Prédiction de phénotypes
- Exploration
- Description

Intégration de données

Points de vue complémentaires :

- Classification (non supervisée, supervisée)
- Sélection de variables, identifications de biomarqueurs
- Prédiction de phénotypes
- Exploration
- Description
- Méta-analyse
- Réseaux
- ...

L'approche statistique est différente selon le point d'entrée adopté.
Ici, nous choisirons la sélection de variables en priorité.

Intégration de données

Problèmes statistiques associés :

- grande dimension (nombre de variables largement supérieur au nombre d'individus)

Intégration de données

Problèmes statistiques associés :

- grande dimension (nombre de variables largement supérieur au nombre d'individus)
- données manquantes
- nature des données hétérogène (qualitatives, quantitatives,...)

Intégration de données

Problèmes statistiques associés :

- grande dimension (nombre de variables largement supérieur au nombre d'individus)
- données manquantes
- nature des données hétérogène (qualitatives, quantitatives,...)
- besoin de normalisation, standardisation (biais techniques intra technologie, unités de mesures différentes, correspondance des variables, ...)
- choix des paramètres, par exemple nombre de groupes, poids pour gérer une qualité des données potentiellement hétérogène

Intégration de données

Problèmes statistiques associés :

- grande dimension (nombre de variables largement supérieur au nombre d'individus)
- données manquantes
- nature des données hétérogène (qualitatives, quantitatives,...)
- besoin de normalisation, standardisation (biais techniques intra technologie, unités de mesures différentes, correspondance des variables, ...)
- choix des paramètres, par exemple nombre de groupes, poids pour gérer une qualité des données potentiellement hétérogène
- besoin d'algorithmes rapides et demandant un espace mémoire raisonnable $\Rightarrow$ quels calculs stocker ? compromis temps de calcul précision ?
- ...

Introduction

Données "omiques" :

- génomiques (ADN)
- transcriptomiques (ARN)
- protéomiques (protéines)
- métabolomiques (métabolites)
- ...

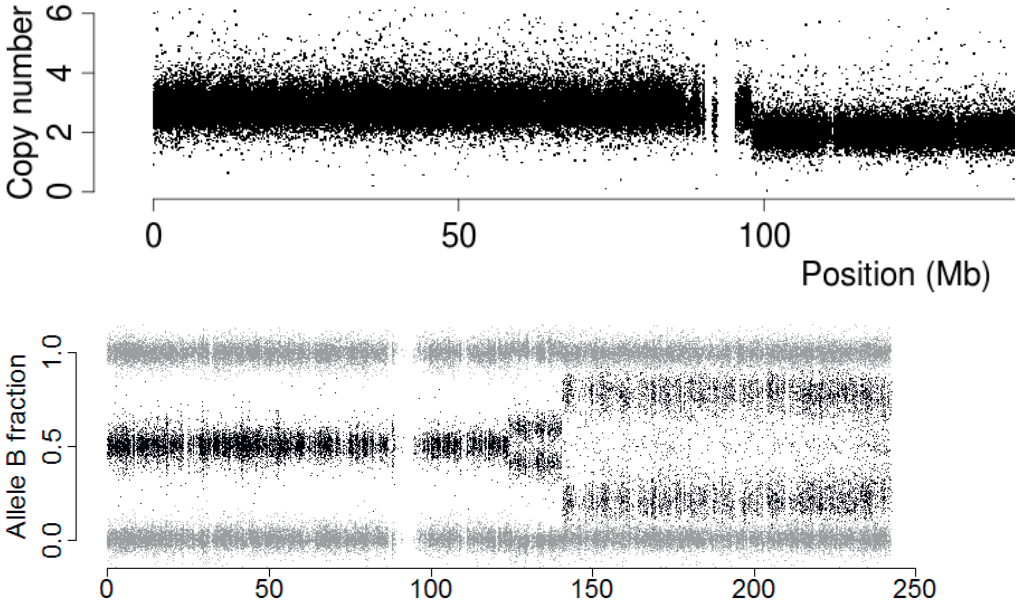
Il existe de plus en plus de bases de données publiques, avec un accès plus ou moins restreint selon le niveau d'analyse (données brutes ou prétraitées)

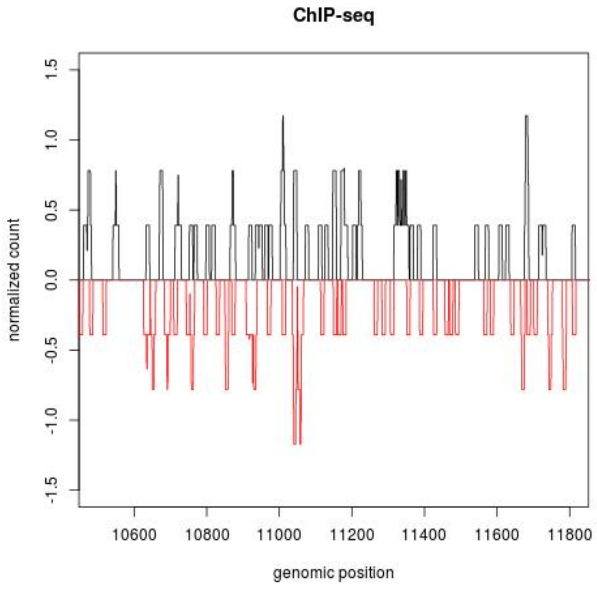
Intégration de données

- en positionnant le long du génome
- sans prendre en compte la position génomique

Intégration de plusieurs jeux de données

Types de signaux à intégrer le long du génome





Piste de recherche : segmentation à noyaux pour capter des ruptures dans la distribution (pas seulement dans la moyenne)

Intégration de plusieurs jeux de données

Jeux de données avec un nombre de variables très différents à intégrer : il vaut mieux ne pas prendre en compte la position le long du génome.

Exemple de méthode d'intégration testée : package SGCCA (A. Tenenhaus, V. Guillemot et al., 2014)

Sparse Generalised Canonical Correlation Analysis

$$\underset{a_1,a_2,\ldots,a_J}{\operatorname{argmax}} \sum_{j \neq k} c_{jk}\, g(\operatorname{cov}(X_j a_j, X_k a_k))$$

$$\text{sous contraintes} \begin{cases} \|a_j\|_2^2 = 1, \ j = 1 \ldots J \\ \|a_j\|_1 \leq c_j, \ j = 1 \ldots J \end{cases}$$

$$\text{avec } c_{jk} = \begin{cases} 1 & \text{si les blocs sont connectés} \\ 0 & \text{sinon} \end{cases}$$

Application sur les données TCGA

Données TCGA (The Cancer Genome Atlas Consortium)

blocs	caractéristiques
X : données d'expression	16653
Z : microARN	470
Y : groupe cytogénétique	3

Test de différentes matrices de design (Siddharth Sharma)

	X	Z	Y
X	0	a	1
Z	a	0	1
Y	1	1	0

Validation croisée pour choisir les paramètres de sparsité

Application sur les données TCGA

a	Nb gènes sélectionnés	Nb miARN sélectionnés
0	1208	438
0.3	306	8
0.5	301	7
0.7	314	7
1	335	7

Résultats :

- Très différents quand $a = 0$ vs $a \neq 0$
- 177 gènes sélectionnés quelque soit la matrice de design
- 6 miARN sélectionnés quelque soit la matrice de design "hsa-miR-136" "hsa-miR-154" "hsa-miR-224" "hsa-miR-376c" "hsa-miR-411" "hsa-miR-758"

Conclusion - perspectives

Conclusion

La question posée reste primordiale à toute analyse $\Rightarrow$ traduire la question biologique sous forme mathématique est un art aussi important que résoudre le problème statistique ainsi posé

Perspectives

L'intégration de données cliniques et génomiques ouvre de belles perspectives de recherche pour les médecins... mais aussi pour les statisticiens...