Complémentarité biologique des méthodes de sélection de gènes

K-A. Lê Cao[1,2], S. Gadat[2], O. Gonçalves, A. Bonnet[1], C. Robert-Granié[1], P. Besse[2]

[1]INRA et [2]Université Paul Sabatier, Toulouse

GDR Statistique et Santé, 2007

Données de microarray

- 1 spot = 1 gène
- Mesure l'expression de gènes dans différentes conditions biologiques
- Mesure de l'expression : intensité du signal
- Spots : allumés ou éteints

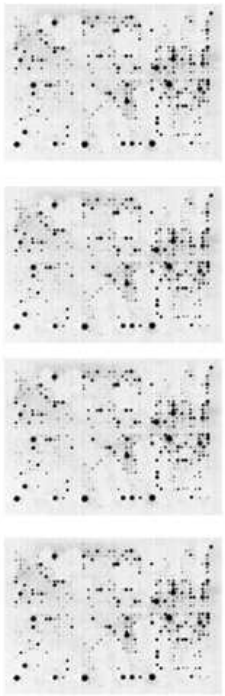


Figure: Normal

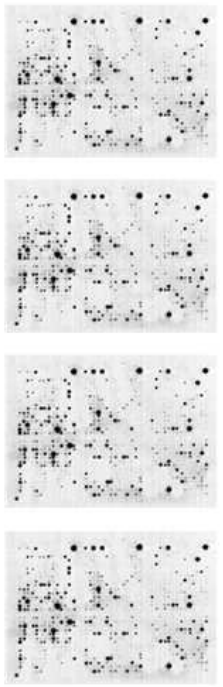


Figure: Cancer

Quels sont les gènes activés, hyper activés, silencieux dans des cellules normales ou cancéreuses ?

Challenge pour les statisticiens . . .

1. Grand nombre de variables p (les gènes), $p > 5000$
2. Petit nombre d'observation n (les microarrays), $n < 50$

. . . et les biologistes

1. plupart des gènes non informatifs (bruit)
2. identifier les gènes régulés lors de l'expérience

$\rightarrow$ *Réduire* le nombre de variables pour inférer de meilleurs résultats (sélection)
$\rightarrow$ Collaboration nécessaire pour analyser et interpréter les résultats

Méthodes filtre

→ significativité d'un gène

→ ordonne selon les p-valeurs (Student, Fisher)

1. résiste au sur-apprentissage
2. faible coût de calcul
3. interactions négligées

Méthodes globales

→ mesure la pertinence d'un sous-ensemble de gènes
→ recherche heuristique ou stochastique

1. risque de sur-apprentissage
2. coût de calcul
3. beaucoup de variables → problème !

Motivation-Contexte

Quels sont les gènes informatifs qui permettent de classer les microarrays ?

Mots clés : classification, selection de variable, problèmes multi-classes déséquilibrés, théorie d'apprentissage

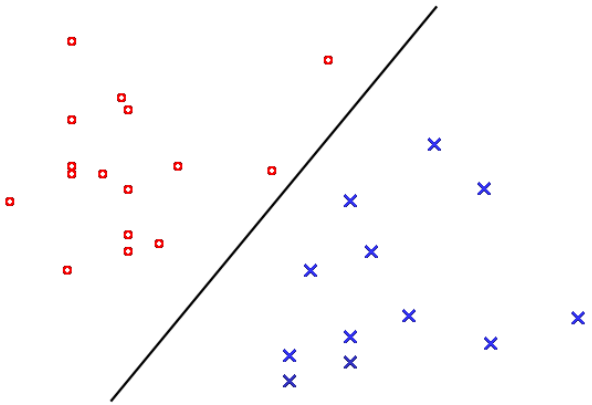
But:

- Présenter une méthode originale de sélection de gène autre que les tests statistiques
- Montrer que les résultats n'ont de valeur que s'ils sont biologiquement interprétés

SVM-CART

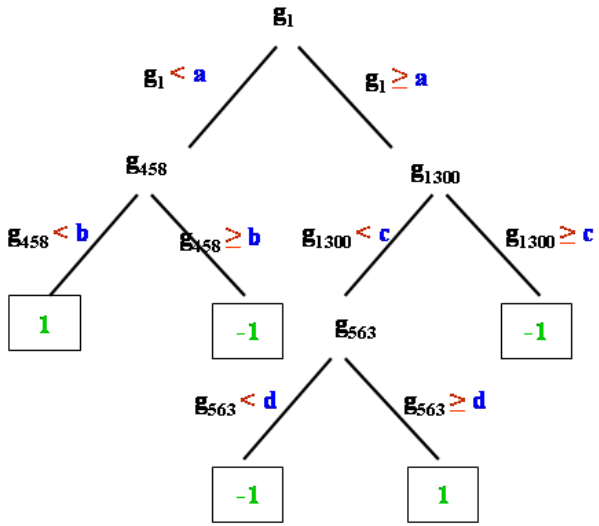
SVM

Construire l'hyperplan optimal qui sépare les classes

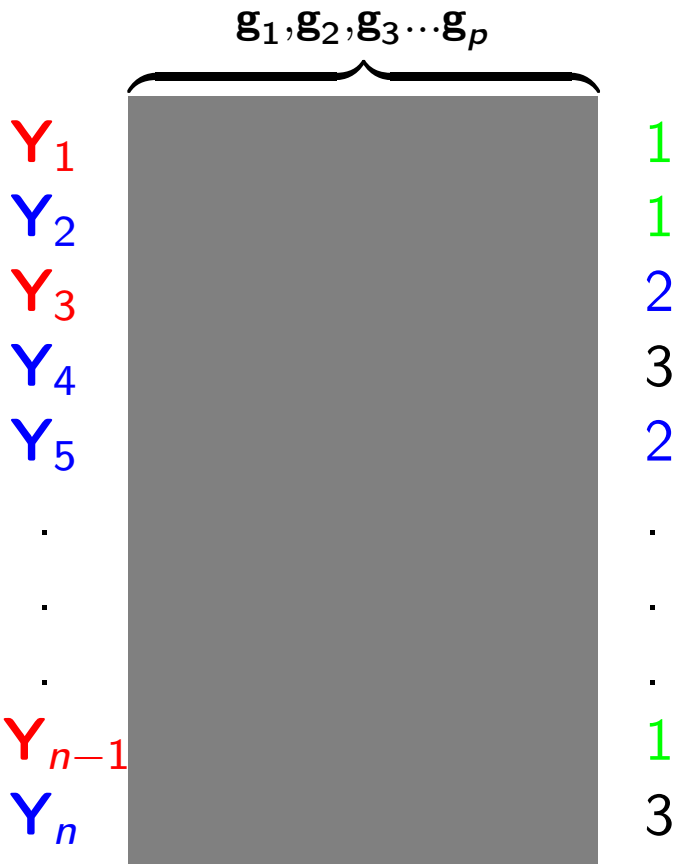


CART

Trouver la meilleure division sur un gène pour chaque noeud



Méthodes d'apprentissage statistique

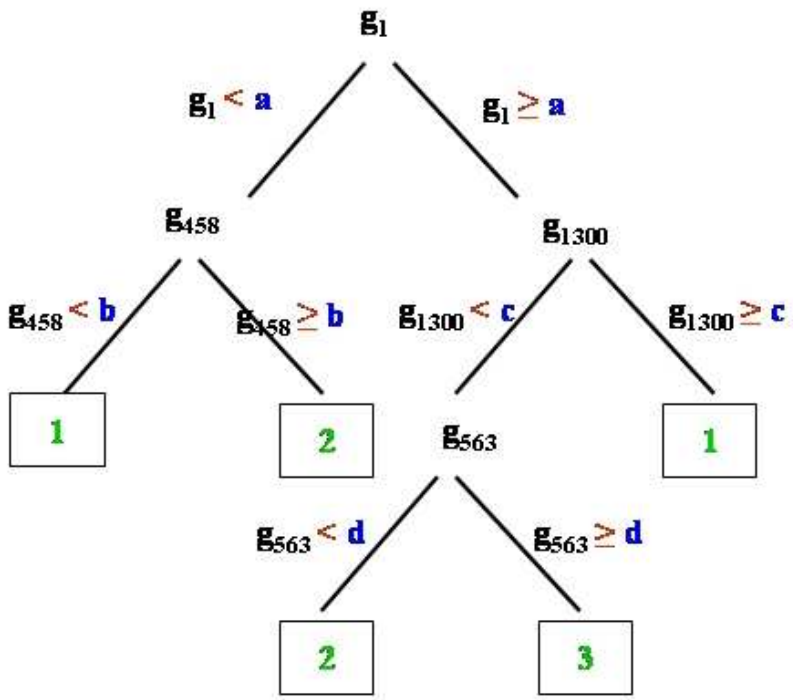


- **But**: apprendre les relations entre les variables et un concept cible spécifique (1 2 3)
  $\rightarrow$ trouver les caractéristiques des variables-gènes
- Généraliser à des exemples non connus
- Echantillon d'apprentissage/test

Phase d'apprentissage

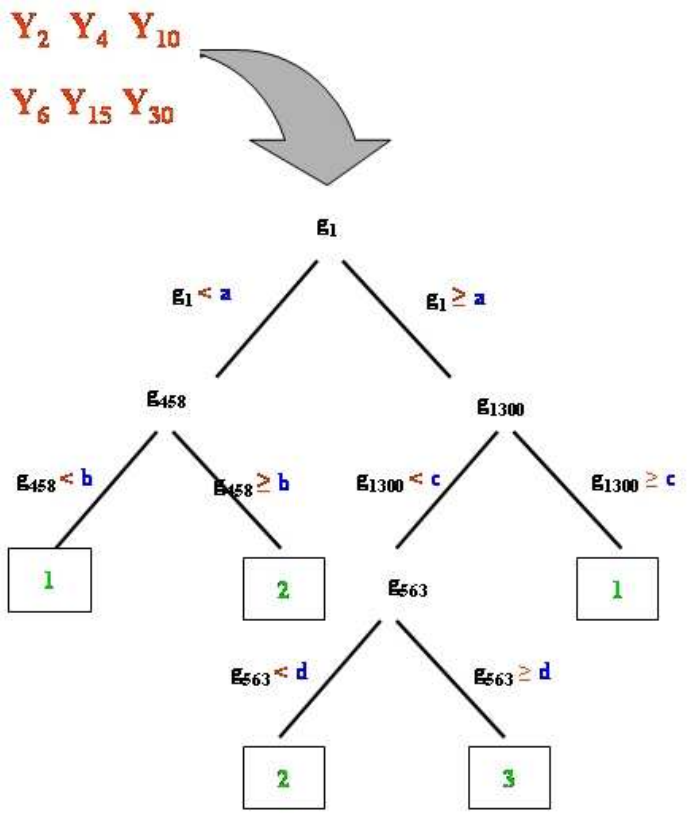
Construire l'arbre sur des données d'apprentissage

- Sélection de la meilleure division (gène + seuil)
- Déclaration du noeud terminal
- Affectation de la classe de chaque noeud terminal



Phase de test

- Faire parcourir les observations test le long de l'arbre
- Comparer leur classe prédite à leur vraie classe
  → Taux d'erreur



Sélection de variables ?

- CART est très instable
- CART ne fait pas de la sélection de variables
- Les méthodes de classification ne font pas de la sélection de variable sauf méthodes wrapper

→ OFW (Optimal Feature Weighting): Algorithme stochastique

1. Sélectionne les gènes pertinents prédicteurs
2. Méthodes de classification appliquées : SVM ou CART
3. Données multi-classes très déséquilibrées et p grand
4. Evaluer la performance statistique avec n petit

Optimal Feature Weighting algorithm (OFW)

Gadat S., Younes L. (2007) A stochastic algorithm for feature selection in pattern recognition, *JMLR* **8**

Notations:
$\mathcal{G}$ ensemble de gènes
ω un sous ensemble de $\mathcal{G}$ (p-uple)
$\mathbb{P}$ une probabilité sur $\mathcal{G}$ (= poids sur chaque gène)

Idée principale:

- estimer la probabilité $\mathbb{P}$ pour qu'elle corresponde à la pertinence de chaque gène dans un cadre de classification
- poids importants = genes discriminants

Fonction de coût

Definition:
ω gènes tirés selon $\mathbb{P}$
$g(\omega)$ mesure l'efficacité de $\omega \in \mathcal{G}^p$ pour classer

$$\varepsilon(\mathbb{P}) = \mathbb{E}_{\mathbb{P}} g(\omega) = \sum_{\omega \in \mathcal{G}^p} g(\omega)\mathbb{P}(\omega)$$

$g(\omega)$ estime l'erreur du classifieur $\mathbb{A}$ construit avec ω.
But : minimiser ε sur $\mathbb{P}$
$\longrightarrow$ algorithme de descente de gradient

Gradient descent

Equation de descente de gradient:

$$\mathbb{P}_{n+1} = \pi[\mathbb{P}_n - \alpha_n \nabla\varepsilon(\mathbb{P}_n)]$$

with

$$\nabla\varepsilon(\mathbb{P})(\delta) = \frac{\partial\varepsilon}{\partial\mathbb{P}(\delta)} = \sum_{\omega\in\mathcal{G}^p} \frac{C(\omega,\delta)\mathbb{P}_p(\omega)}{\mathbb{P}(\delta)} g(\omega) \qquad (1)$$

où δ est un gène dans ω et $C(\omega,\delta)$ le nombre d'occurences de δ dans ω
$\rightarrow$ insoluble !

$\rightarrow$ remplacer la somme de tous les $\omega \in \mathcal{G}^p$ possibles par un ω_n:

$$d_n = \frac{C(\omega_n,\cdot)g(\omega_n)}{\mathbb{P}_n(\cdot)}$$

Algorithme détaillé

Pour $n = 0$, $\mathbb{P}_0$ est uniforme sur $\mathcal{G}$
Tant que $||\mathbb{P}_{n+k} - \mathbb{P}_n|| > C$:

- extraire ω_n de $\mathcal{G}^p$ suivant $\mathbb{P}_n^{\otimes p}$
- construire le classifieur avec ω_n
- calculer le taux d'erreur $g(\omega_n)$ et le vecteur d_n
- mettre à jour $\mathbb{P}_{n+1} = \pi[\mathbb{P}_n - \alpha_n d_n]$
- $n = n + 1$

Calcul du gradient

ofw+SVM

- construire 1 SVM_{ω_n} sur 1 échantillon bootstrap
- calculer le taux d'erreur sur échantillon OOB

ofw+CART (instable $\rightarrow$: aggréger)

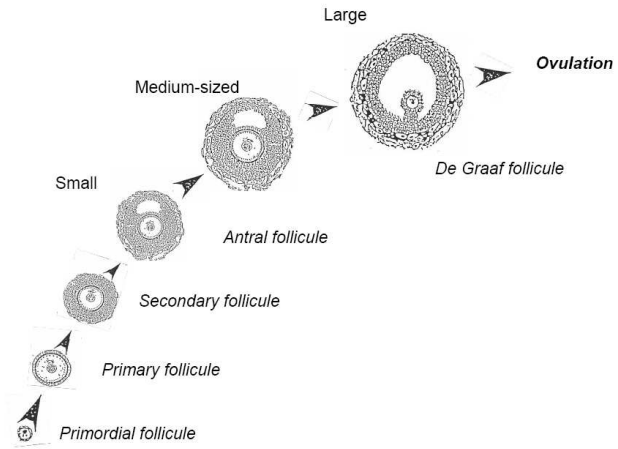
- construire k $CART_{\omega_n^k}$ sur k échantillons bootstrap
- calculer le taux d'erreur moyen $\bar{g}(\omega_n)$ sur les k échantillons OOB

Jeux de données

data set	Colon	Brain	Folliculogénèse	Eadgene
# genes	2000	1963	1564	7000
# classes	2 N/C	5	3	4 time points
# obs	62= 40 + 22	42=10 + 10 + 10+4+8	42=20 + 14+8	15=4 + 4 + 3+ 4

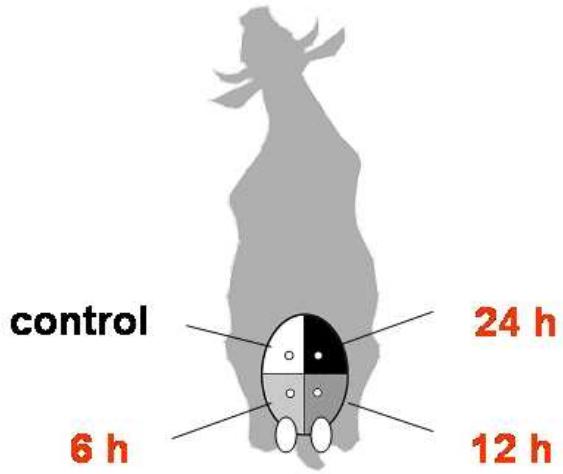
Les jeux de données sont correctement normalisés
Quelques jeux de données ont été pré-filtrés

Folliculogénèse



→ comprendre les différents mécanismes du développement folliculaire de la truie

Eadgene



→ identifier les gènes dont l'expression varie suite à une infection artificielle avec la bactérie E.Coli

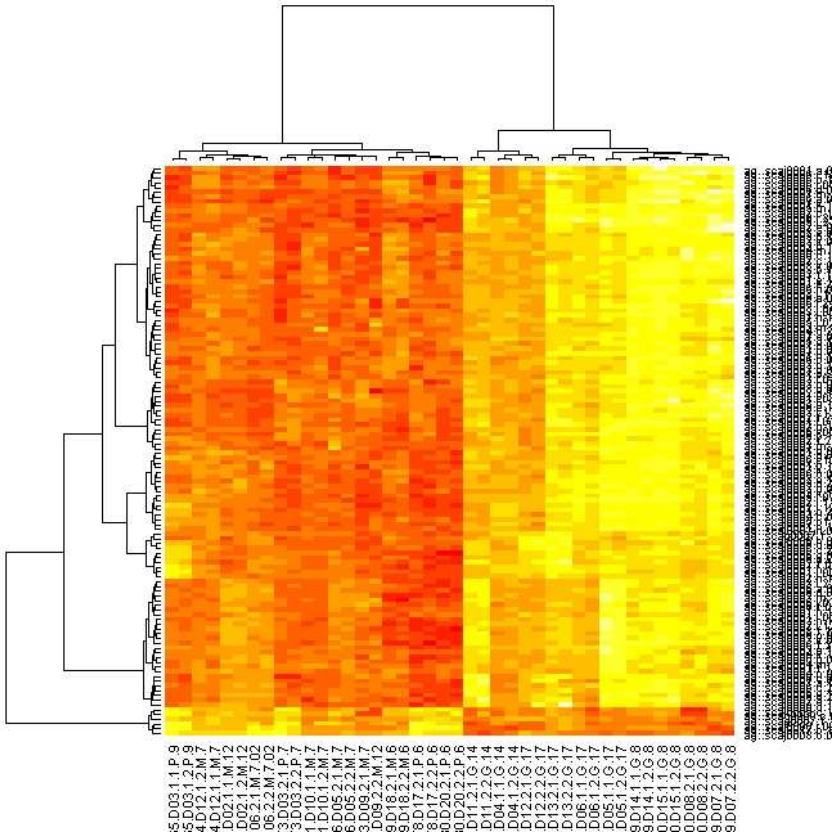


Figure: Sélection F-test

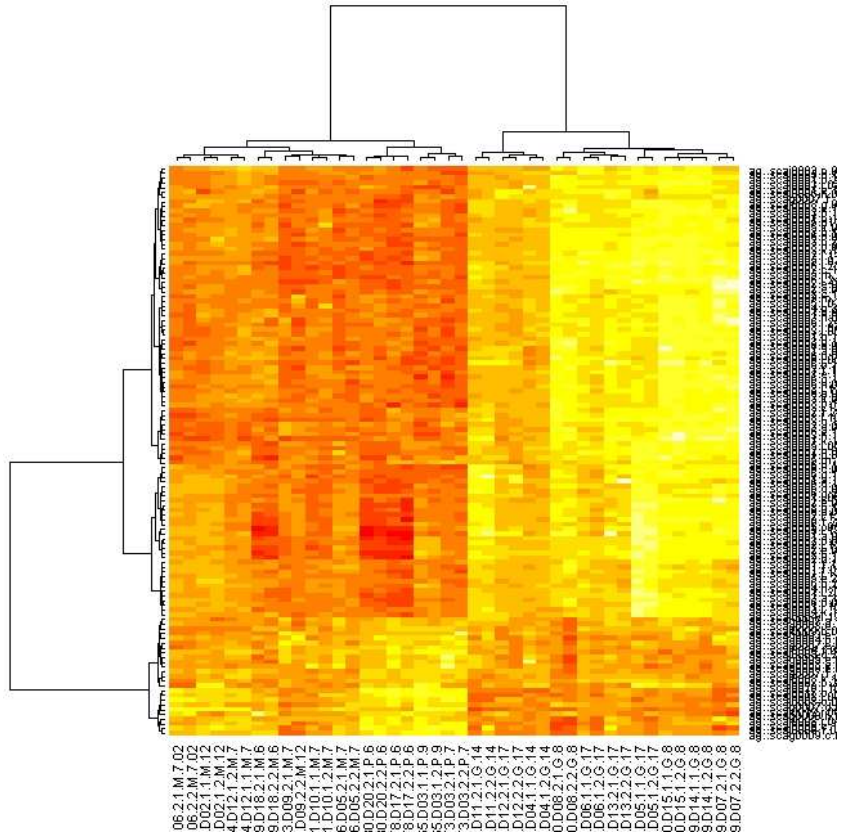
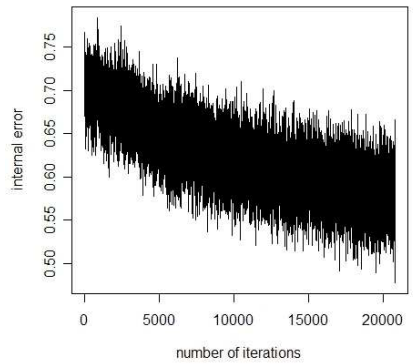
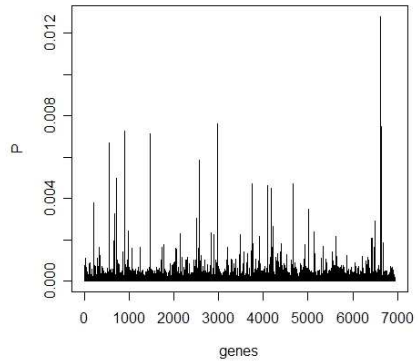


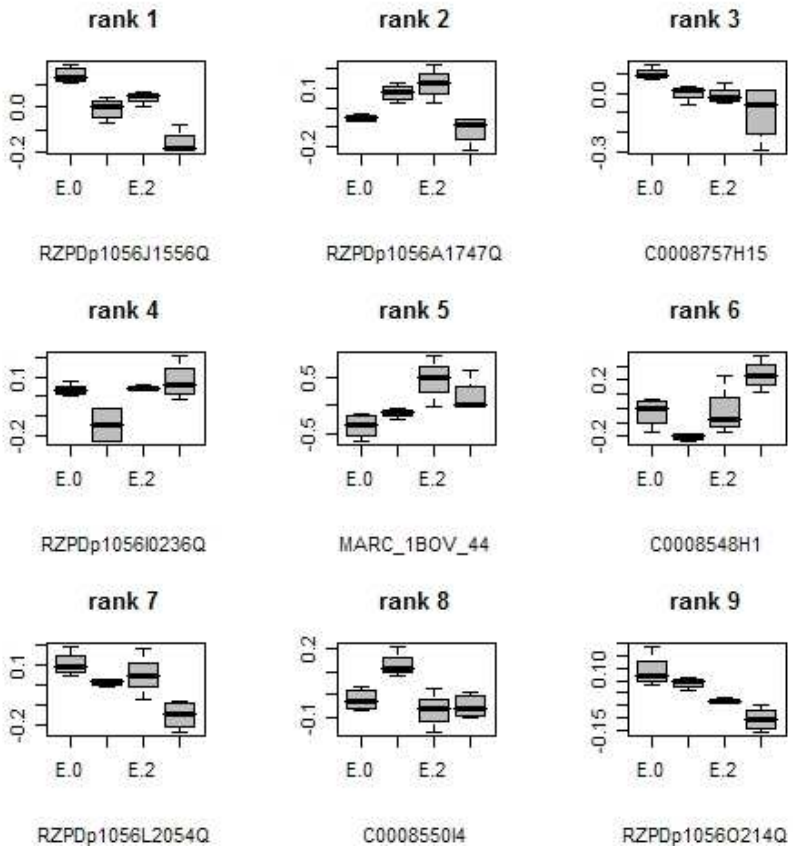
Figure: Sélection Random Forest

"Identification of gene networks involved in pig antral follicular development", Bonnet et al., soumis

Eadgene


internal error
0.50
0.55
0.60
0.65
0.70
0.75
0
5000
10000
15000
20000
number of iterations


P
0.000
0.004
0.008
0.012
0
1000
2000
3000
4000
5000
6000
7000
genes


rank 1
E.0
E.2
RZPDp1056J1556Q
rank 2
RZPDp1056A1747Q
rank 3
C0008757H15
rank 4
RZPDp1056I0236Q
rank 5
MARC_1BOV_44
rank 6
C0008548H1
rank 7
RZPDp1056L2054Q
rank 8
C0008550I4
rank 9
RZPDp1056O214Q

Comparaison avec différentes méthodes

1. Méthodes SVM: RFE (Guyon et al., 2002), l_0 SVM (Weston et al., 2003)
2. Méthodes CART: Random Forests (Breiman, 2001)
3. Méthodes filtre: T-test, F-test

Evaluation avec le taux d'erreur e632+ bootstrap (Efron, 1983)

- Comparer les différentes méthodes
- Evaluer le taux d'erreur même si n petit

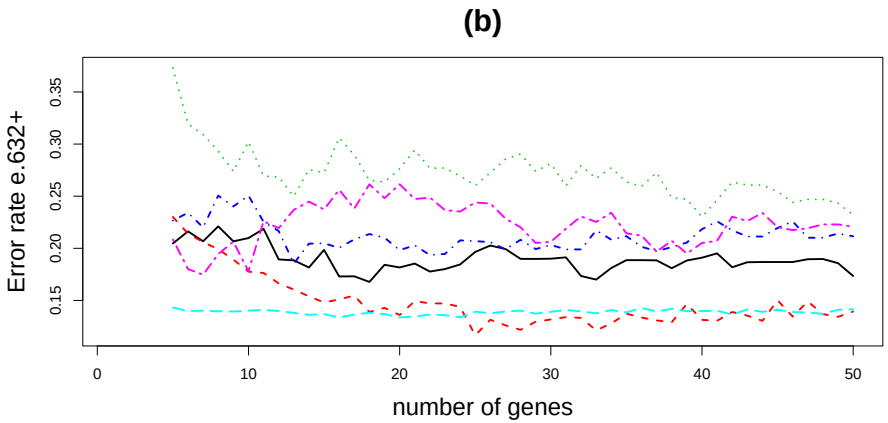


Figure: Colon-2 class

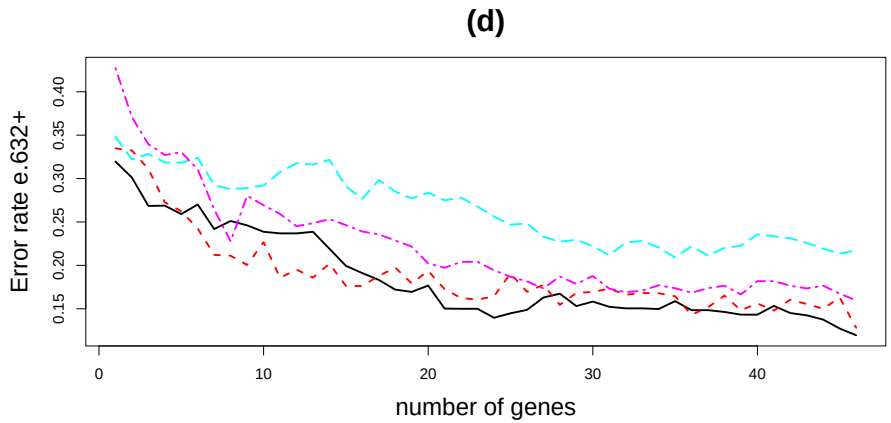
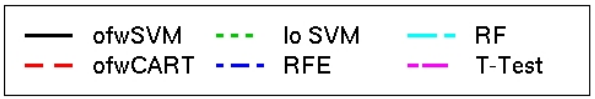


Figure: Brain-5 class



Comparaisons des sélections de gènes

Brain \ Colon	ofwSVM	RFE	l0	ofwCART	RF	T-test
ofwSVM	#	14	13	6	0	5
RFE		#	27	0	0	0
l0			#	1	0	0
ofwCART	8			#	16	16
RF	8			17	#	36
T-test (F-test)	5			4	2	#

Interprétation biologique

Olivier Gonçalves (IUT Clermont-Ferrand)

→Manual curating
→Biologie intégrative avec Ingenuity Pathways Analyses
http://www.ingenuity.com/

- Réseaux de gènes régulés
- Fonction biologiques
- Voies de signalisation

Method / Criterion	T-test	RF	ofw CART	ofw SVM	l_0 SVM	RFE
Number of networks	4	4	6	4	4	5
Cancer term frequency in networks	1	3	2	1	2	1
Gastrointestinal disease term frequency in networks	0	1	0	0	0	2
Rank of the ontological term in the function list:						
Cancer	11	17	6	4	11	15
Gastrointestinal disease	43	67	49	67	0	22
Tissue development	45	NA	36	2	2	2
Tissue morphology	1	1	37	39	35	26
Skeletal and muscular syst. dev.	3	2	35	40	5	6
Number of genes associated with the ontological term:						
Cancer	11	12	8	12	6	8
Tissue development	2	0	3	6	5	7
Tissue morphology	9	11	8	5	8	6
Skeletal and muscular syst. dev.	12	12	7	7	12	9
Colon Cancer	2	1	0	2	0	1
Colon cancer gene name	CDH3 GUCA2B	CDH3		CDH3 GUCA2B		GUCA2B
Genes involved in the signaling pathways:						
PI3K/AKT			Bcl2	PPP2R5		MEF2C
ERK/MAPK				ETS2		PPP2RC
p38/MAPK				PPP2R5	IL1R2, MEF2	MEF2
Wnt/Beta Catenin	CDH3	CDH3	CSNK2A2	CDH3		PPP2R5C

Conclusion

- Sous ensemble de gène optimal pertinent
- Résultats compétitifs avec d'autres méthodes sur données publiques
- Interprétation biologique
- Sélection différentes mettent en valeur différentes relations entre les gènes

"Selection of biologically relevant genes with a stochastic algorithm"

Lê Cao, K-A., Gonçalves, O., Besse, P., Gadat, S., Statistical Applications in Genetics and Molecular Biology: Vol. 6, 2007

- Packages R
  - CART: *rpart*
  - SVM: *e1071*
  - Random Forests : *randomForest*
  - Optimal Feature Weighting: *ofw*...bientôt!
- Généralisation au multiclasse (pondération de l'erreur, soumis)

Stochastic optimization method-Annexe

Idea: find d_n such that

$$\mathbb{E}[d_n] = \nabla_{\mathbb{P}_n} \varepsilon$$

solved using

$$d_n = \frac{C(\omega_n, \cdot) g(\omega_n)}{\mathbb{P}_n(\cdot)}$$

$\longrightarrow$ Robbins-Monro: use a stochastic approximation algorithm that will replace the expectation by only one sample ω_n

Proof of the convergence in (Gadat et al. 2005)