N° d'ordre: 936

Thèse présentée en vue de l'obtention de grade de

Docteur de l'Institut National des Sciences Appliquées de Toulouse

Spécialité : Mathématiques et Applications

par

Kim-Anh Lê Cao

Titre de la thèse :

Outils statistiques pour la sélection de variables et l'intégration de données "omiques"

Soutenue le 1er octobre 2008 devant la commission d'examen

M. :	Alain	Baccini	Président
MM. :	Gilles	Celeux	Rapporteurs
	Geoff	McLachlan	
MM. :	Anestis	Antoniadis	Examinateurs
	Sandrine	Lagarrigue	
MM. :	Philippe	Besse	Directeurs de thèse
	Christèle	Robert-Granié	

Station d'Amélioration Génétique Animale - UR 631, Institut National de la Recherche Agronomique, BP 52627, 31326 Castanet-Tolosan cedex
Institut de Mathématiques, Université de Toulouse et CNRS (UMR 5219), 31062 Toulouse cedex 9

Merci !

Et voilà la page tant attendue, celle qui finalise la thèse et ces trois belles années. L'unique page qui sera lue par plus de 6 personnes, si si, j'ai espoir !
Parfois, pendant la thèse, on peut se sentir seul(e). Mais on a beau dire, on n'est pas seul(e) tout le temps, loin de là. Toutes mes pensées vont à ceux qui m'ont tant aidée, tant apporté pendant ces trois années, et j'aimerais citer :

Mes deux rapporteurs, Gilles Celeux et Geoff McLachlan, pour l'intérêt qu'ils ont apporté à mon manuscrit. *Thank you Geoff for the time we spent together, and for showing me that everything I hoped for could happen.* Je remercie aussi les membres du jury, Anestis Antoniadis et Sandrine Lagarrigue, pour avoir accepté de venir de si loin pour participer à ma soutenance.

Je voudrais remercier l'équipe de "choc" qui m'a encadrée, je nomme Christèle Robert-Granié et Philippe Besse, sans qui rien de tout cela n'aurait pu être possible. Christèle, merci d'avoir toujours été là, si près (la porte à côté en fait), et toujours de bonne humeur. Philippe, merci de m'avoir guidée avec sagesse et justesse. Merci à tous les deux de m'avoir permis de faire une chose à laquelle j'aspirais depuis longtemps, partir faire un séjour à Cape Town. Je vous souhaite à tous les deux d'encadrer beaucoup d'autres petits thésards fougueux et dynamiques.

Je remercie bien sûr toutes les personnes qui ont participé de (très) près à cette thèse. Sébastien Déjean, qui a été mon mentor en `R` depuis que je suis toute petite (enfin, façon de parler bien sûr, vu que je suis restée petite, je te vois venir avec tes remarques). Sébastien Gadat, qui a été très présent pendant la première partie de ma thèse. Merci pour ton optimisme débordant lorsque rien ne semblait marcher comme il fallait, et merci de m'avoir montré le premier qu'un travail appliqué pouvait aussi être du bon travail. Merci à Olivier Gonçalves, qui, le premier, a ouvert de nouvelles perspectives d'analyse en nous présentant Ingenuity. Un très **très** grand merci à Patrick Chabrier, pour le temps que tu as passé avec moi à débugger mes programmes. Tu m'as beaucoup appris, devant un ordi comme sur les voies. Pour ce dernier point, j'espère que nous n'en resterons pas là ! *Thank you Debra Rossouw, for your efficiency, it was a great pleasure working with you.* Merci à Pascal Martin, mon biologiste modèle, celui qui a l'air de manier aussi bien les pipettes que les programmes en `R` pour faire de superbes graphiques. Merci aux filles de Jouy, Isabelle Hue et Séverine Degrelle, pour les nombreux échanges et au final, une belle histoire qui se profile non ? *A big THANK YOU to Elizabeth Kelly for correcting many of my frenglish mistakes. You cannot imagine how relieved I felt during this stressing period.* Merci au joyeux groupe "puces" du LGC : les filles, vous avez été les premières a me montrer que travailler en collaboration pouvait (et devrait !) se faire dans la bonne humeur. Merci à Agnès Bonnet, Gwenola

Tosser, Laurence Liaubet, qui font partie de tous ces biologistes qui ont essayé, à de nombreuses reprises, de m'inculquer quelques notions de biologie. Je ne dis pas que j'ai tout retenu, mais j'espère l'essentiel ! Un grand merci également à Magali San Cristobal, qui m'a montré la voie.
Special thanks to Cathal Seoighes for welcoming me into the UCT Computational Biology group. I did spend a great time there. Merci à Alain Baccini, pour ses sages conseils et le temps qu'il m'a accordé. Merci à Nicolas Renon pour m'avoir aidée à lancer mes calculs avec CALMIP.
Merci bien sûr à la SAGA pour son accueil, ses pots et ses gâteaux à la salle café. Merci à Valérie, Dounia, Nancy, Line et Carine pour leur efficacité à toute épreuve. Je n'oublie pas Mylène, Isabelle, Virginie (j'ai eu du mal à me remettre de votre surprise-partie, je suis une émotive moi), Anne, toujours prête à envoyer plein de mails le matin (oups, j'ai fait une gaffe là ?) et Andrés.
Je remercie également les enseignants chercheurs du GMM de l'INSA qui accompagnent et guident les jeunes moniteurs.

Et puis à côté de la thèse, il y a le reste, et là ça devient moins sérieux.

Je remercie tout le petit groupe de l'Adas Escalade, qui, grâce à nos escapades tri-hebdomadaires à Alti, m'a permis de m'aérer le cerveau et me faire les muscles. Merci en particulier à Matthias, le daltonien qui confond gauche et droite, Cécé qui passe tout droit dans les rond-points et le multi-sportif Jérôme, toujours à bloc, aren't you ? Simon, compagnon de grimpe, de congrès, de restau et autres, l'homme parfait sanitairement parlant, j'espère que tu viendras nous voir dans l'hémisphère sud (tu as le droit de choisir entre deux continents). Merci à feu la croisière s'amuse, avec qui tout a commencé pour de bon, Rika et Nono, merci de m'avoir prêté votre lit, un soir difficile, Rom (c'est fini les bérouz) et Mèl, merci pour les soirées rouges et fromages et le coup des poubelles, Carrère et ses messages intempestifs, Soso, Gégé et Lulu. Merci aussi à Nicklas, the cocktail party man. Merci à Iadine et Tara pour m'avoir dicté le bon choix à faire et Lolo. Merci à tout ce beau monde avec qui j'ai beaucoup rigolé et un peu picolé (c'est pour la rime bien sûr).
Merci à ceux que je n'ai pas eu le temps de trop croiser pendant cette période, pour cause d'éloignement kilométrique (petit JB, Blackos, Piju), ou pas (l'Olive, Mathieu, Thierry, merci pour les ciné). Merci Marion de m'avoir initiée à la beauté du cirque.

At last, but not least, je remercie mes plus proches. Mes parents qui m'ont offert tant de possibilités ... et tant de voyages. Merci mon petit père pour toujours t'être inquiété de "ma thèse", merci ma petite mère pour ta sérénité (il est loin le temps où j'avais "perdu ma logique"). J'espère que nous aurons l'occasion encore et toujours, de se retrouver à l'autre bout du monde, pour manger du gorgonzola à la cuillère (il faudra bien s'y mettre) ou faire du triple galop dans le *free state*. Merci à mon petit frère, qui voit la vie en dimension rectangulaire en noir et blanc avec un talent inimitable. Merci pour les moments complices passés. Enfin, merci à Olivier, my beloved, qui a toujours cru en moi, et qui a toujours été là, même de loin.

A mes parents,
A mon frère Vinh

This is a film about a man and a fish
This is a film about dramatic relationship between man and fish
The man stands between life and death
The man thinks
The horse thinks
The sheep thinks
The cow thinks
The dog thinks
The fish doesn't think
The fish is mute, expressionless
The fish doesn't think because the fish knows everything
The fish knows everything
Arizona Dream, Goran Bregovic

Résumé

Les récentes avancées biotechnologiques permettent maintenant de mesurer une énorme quantité de données biologiques de différentes sources (données génomiques, protéomiques, métabolomiques, phénotypiques), souvent caractérisées par un petit nombre d'échantillons ou d'observations.

L'objectif de ce travail est de développer ou d'adapter des méthodes statistiques adéquates permettant d'analyser ces jeux de données de grande dimension, en proposant aux biologistes des outils efficaces pour sélectionner les variables les plus pertinentes. Dans un premier temps, nous nous intéressons spécifiquement aux données de transcriptome et à la sélection de gènes discriminants dans un cadre de classification supervisée. Puis, dans un autre contexte, nous cherchons à sélectionner des variables de types différents lors de la réconciliation (ou l'intégration) de deux tableaux de données omiques.

Dans la première partie de ce travail, nous proposons une approche de type *wrapper* en agrégeant des méthodes de classification (CART, SVM) pour sélectionner des gènes discriminants une ou plusieurs conditions biologiques. Dans la deuxième partie, nous développons une approche PLS avec pénalisation l_1 dite de type *sparse* car conduisant à un ensemble "creux" de paramètres, permettant de sélectionner des sous-ensembles de variables conjointement mesurées sur les mêmes échantillons biologiques. Un cadre de régression, ou d'analyse canonique est proposé pour répondre spécifiquement à la question biologique.

Nous évaluons chacune des approches proposées en les comparant sur de nombreux jeux de données réels à des méthodes similaires proposées dans la littérature. Les critères statistiques usuels que nous appliquons sont souvent limités par le petit nombre d'échantillons. Par conséquent, nous nous efforçons de toujours combiner nos évaluations statistiques avec une interprétation biologique détaillée des résultats.

Les approches que nous proposons sont facilement applicables et donnent des résultats très satisfaisants qui répondent aux attentes des biologistes.

Mots clés : sélection de variables, classification, sparse PLS, algorithme stochastique, biologie intégrative.

Statistical tools for variable selection and integration of omics data

Abstract

Recent advances in biotechnology allow the monitoring of large quantities of biological data of various types, such as genomics, proteomics, metabolomics, phenotypes..., that are often characterized by a small number of samples or observations.

The aim of this thesis was to develop, or adapt, appropriate statistical methodologies to analyse highly dimensional data, and to present efficient tools to biologists for selecting the most biologically relevant variables. In the first part, we focus on microarray data in a classification framework, and on the selection of discriminative genes. In the second part, in the context of data integration, we focus on the selection of different types of variables with two-block omics data.

Firstly, we propose a *wrapper* method, which agregates two classifiers (CART or SVM) to select discriminative genes for binary or multiclass biological conditions. Secondly, we develop a PLS variant called *sparse PLS* that adapts l_1 penalization and allows for the selection of a subset of variables, which are measured from the same biological samples. Either a regression or canonical analysis frameworks are proposed to answer biological questions correctly.

We assess each of the proposed approaches by comparing them to similar methods known in the literature on numerous real data sets. The statistical criteria that we use are often limited by the small number of samples. We always try, therefore, to combine statistical assessments with a thorough biological interpretation of the results.

The approaches that we propose are easy to apply and give relevant results that answer the biologists needs.

Keywords : feature selection, classification, sparse PLS, stochastic algorithm, systems biology.

Liste des travaux

Degrelle S. and Lê Cao K.-A., Robert-Granié C., Hue I. (2008) Molecular prediction of gastrulation stages stages with a small extra-embryonic gene-set (*en préparation*).

Lê Cao K.-A. and Martin P.G.P., Robert-Granié C., Besse P. (2008) Sparse canonical methods for biological data integration : application to a cross-platform study (*soumis*).

Lê Cao K.-A., Rossouw D., Robert-Granié C., Besse P. (2008) Sparse PLS : Variable Selection when Integrating Omics data (*soumis*).

Lê Cao K.-A., Chabrier P. (2008) ofw : an R package to select continuous variables for multiclass classification with a stochastic wrapper method. Journal of Statistical Software (*sous presse*).

Lê Cao K.-A., Bonnet A., Gadat S. (2007) Multiclass classification and gene selection with a stochastic algorithm (*soumis*).

Tosser-Klopp G., Lê Cao K.-A., Bonnet A., Gobert N., Hatey F., Robert-Granié C., Déjean S., Antic J., Baschet L., San Cristobal M. (2008) A pilot study on transcriptome data analysis of folliculogenesis in pigs. Animal (*sous presse*, non inclus dans ce document).

Bonnet A., Lê Cao K.-A., Sancristobal M., Benne F., Robert-Granié C., Law-So G., Fabre S., Besse P., De Billy E., Quesnel H., Hatey F., Tosser-Klopp G. (2008) Identification of gene networks involved in antral follicular development. Reproduction, 136, 211-224.

Lê Cao K.-A., Gonçalves 0., Besse P., Gadat S. (2007) Selection of Biologically Relevant Genes with a Wrapper Stochastic Algorithm. Statistical Applications in Genetics and Molecular Biology, Vol. 6 : Iss. 1, Article 29.

Contributions au groupe de travail EADGENE[1] (non inclus dans ce document) :

De Koning D.J., Jaffrézic F., Sandø Lund M., Watson M., Channing C., Hulsegge I., Pool M.H., Buitenhuis B., Hedegaard J., Hornshøj H., Jiang L., Sørensen P., Marot G., Delmas C., Lê Cao K.-A., San Cristobal M., Baron M.D., Malinverni R., Stella A., Brunner R.M., Seyfert H.-M., Jensen K., Mouzaki D., Waddington D., Jiménez-Marík A., Pérez-Alegre M., Pérez-Reinado E., Closset R., Detilleux J.C., Dovc P., Lavric M., Nie H. and Janss L. (2007) The EADGENE Microarray Data Analysis Workshop. Genetics Selection Evolution, 39, 621-633.

Jaffrézic F., De Koning D.J., Boettcher P.J., Bonnet A., Buitenhuis B., Closset R., Déjean S., Delmas C., Detilleux J.C., Dovc P., Duval M., Foulley J.L., Hedegaard J., Hornshøj H., Hulsegge I.B., Janss L., Jensen K., Jiang L., Lavric M., Lê Cao K.-A., Sandø Lund M., Malinverni R., Marot G., Nie H., Petz M., Pool M.H., Robert-Granié C., San Cristobal M., van Schothorst E.M., Schuberth H.J., Sørensen P., Stella A., Tosser-Kopp G., Waddington D., Watson M., Yang W., Zerbe H., Seyfert H.-M. (2007) Analysis of the real EADGENE data set : Comparison of methods and guidelines for data normalization and selection of differentially expressed genes. Genetics Selection Evolution, 39, 633-650.

Sørensen P, Bonnet A, Buitenhuis B, Closset R, Déjean S, Delmas C., Duval M, Glass L, Hedegaard J, Hornshøj H, Hulsegge IB, Jaffrezic F, Jensen K, Jiang L, de Koning DJ, Lê Cao K.-A., Nie H, Petz W, Pool MH, Robert-Granie C., SanCristobal M, Sandø Lund M, van Schothorst EM, Schuberth HJ, Seyfert HM, Tosser-Klopp G, Waddington D, Watson M, Yang W, Zerbe H (2007) Analysis of the real EADGENE data set : Multivariate approaches and post analysis. Genetics Selection Evolution, 39, 651-668.

Watson M., Pérez Alegre M., Baron M., Delmas C., Dovc P., Duval M., Foulley J.L., Garrido-Pavón J.J., Hulsegge B., Jaffrézic F., Jiménez-Marín A., Lavric M., Lê Cao K.-A., Marot G., Pool M.H., Robert-Granie C., San Cristobal M., Tosser-Klopp G., Waddington D., De Koning D.J. (2007) Analysis of a simulated microarray dataset : Comparison of methods for data normalization and detection of differential expression. Genetics Selection Evolution,39, 669-689.

[1] European Animal Disease Genomics Network of Excellence for Animal Health and Food Safety, `http://eadgene.info/NewsandEvents/EADGENEEvents/DataAnalysisWorkshopNov2006/tabid/166/Default.aspx`

Table des matières

1. Introduction

Depuis une dizaine d'années, les développements en biotechnologie ont permis à la biologie moléculaire de mesurer l'information contenue dans des milliers de gènes grâce aux puces à ADN (ou *microarray* en anglais). Ceci a permis de déterminer les gènes exprimés dans une condition donnée, et ainsi de mieux comprendre certains processus biologiques mis en œuvre lors de l'expérience. Plus récemment, le développement d'autres supports a permis de mesurer l'expression d'autres types de données, telles que les données issues du protéome, ou du métabolome. Dans ce contexte de biologie systémique, le but est d'identifier des relations entre ces données issues de niveaux fonctionnels différents et de comprendre les interactions souvent complexes entre ces différents composants.

Le volume et la spécificité des jeux de données ainsi constitués conduisent à des traitements faisant appel à tout l'éventail méthodologique statistique, tels que l'exploration de données (Analyse en Composantes Principales, classification non supervisée), les tests statistiques pour déterminer les gènes différentiellement exprimés d'une condition biologique à une autre, la modélisation (modèle linéaire, modèle mixte), ou l'apprentissage (Classification and Regression Trees (CART), Support Vector Machines (SVM), agrégation de modèles), afin d'extraire des informations pertinentes pour le biologiste.
L'objectif de ce travail est d'analyser des situations complexes et d'y adapter les méthodes statistiques adéquates, dans un premier temps pour sélectionner des variables dans le cadre de données d'expression et de classification supervisée, puis dans un deuxième temps pour développer des méthodes pour la mise en correspondance de données d'expression génomiques et d'autres variables omiques (protéomiques, métabolomiques) ou phénotypiques conjointement observées.

Le problème spécifique de la sélection de variables nécessite une approche particulière puisque le nombre de variables est très largement supérieur au nombre d'expériences (ou d'observations). Les approches envisagées relèvent des méthodes de classification (CART, SVM) ainsi que des méthodes de type *wrapper* ou *embedded*, ces méthodes sélectionnant de façon implicite les variables (*Random Forests*, Breiman, 2001 ; *Recursive Feature Elimination*, Guyon *et al.*, 2002) à l'opposé des méthodes *filtres*, approche couramment utilisée à ce jour pour analyser des données du transcriptome. Une fois les gènes actifs repérés, il est essentiel de donner une interprétation biologique afin de mesurer leur pertinence par rapport à la problématique biologique étudiée, d'autant plus que, compte tenu du petit nombre d'échantillons, les critères statistiques

pour évaluer de telles méthodes sont très limités. Un travail de collaboration entre les deux disciplines (statistique et biologie) est donc nécessaire pour valider les approches proposées.

Dans le cadre d'intégration de données de type omiques, une stratégie différente de celle proposée ci-dessus doit être adoptée, afin de mettre en relation des données du transcriptome, avec des données protéomiques ou métabolomiques mesurées sur les mêmes individus. Ici nous nous plaçons dans le cadre de ce que l'on appelle la "réconciliation des données", qui est observée à différents niveaux. Il faut pour cela envisager des approches exploratoires telles que l'analyse en composantes principales et des méthodes de régression multivariée comme l'analyse des corrélations canoniques (CCA) ou la régression PLS (Partial Least Squares). Le but est d'appliquer une pénalisation de type lasso (Tibshirani, 1996) sur ces méthodes de régression, afin d'obtenir des méthodes parcimonieuses (appelées *sparse* en anglais) et de sélectionner dans chaque groupe les variables les plus pertinentes. Encore une fois, la validation de telles approches nécessite une contribution importante de la part du biologiste, dont le travail d'interprétation doit être facilité à l'aide de représentations graphiques appropriées.

Notre objectif est donc d'adapter et de développer les outils statistiques rendus indispensables par l'approche "intégrative" de la biologie, et de mettre en valeur le travail de collaboration nécessaire entre statisticiens et biologistes pour interpréter les résultats de ces nouveaux outils, sur des données souvent complexes.

Ce travail de thèse est composé de trois parties :

1. La première partie traite le problème de la sélection de gènes, dans le cadre de la classification supervisée. Le contexte particulier de sélection de variables pour les données du transcriptome est présenté. Nous développons ensuite trois travaux. Le premier présente une méthode de sélection de gènes, *Optimal Feature Weighting* (OFW, Gadat & Younes, 2007), implémentée avec deux types de classifieurs, SVM ou CART, et développée dans un cadre de classification binaire sur des données publiques de cancer. Nous comparons les résultats obtenus avec d'autres méthodes de types *wrapper, embedded* et filtre selon des critères statistiques, mais surtout biologiques. Le deuxième travail se place dans la continuité du premier article, en s'intéressant cette fois au cas multiclasse déséquilibré, caractéristique des données du transcriptome. Plusieurs critères statistiques sont proposés pour évaluer les différentes méthodes considerées, ainsi qu'une interprétation biologique préliminaire concernant les deux variantes de l'algorithme proposé. Enfin le troisième travail a pour but de valoriser notre approche en facilitant son usage grâce au logiciel `R`.
   Nous concluons cette première partie par un bilan et des perspectives de travail futur.

2. La deuxième partie traite du problème de sélection de variables dans le cadre d'intégration de données omiques. Plusieurs approches parcimonieuses de méthodes d'exploration de données et de méthodes d'intégration de données issues de deux tableaux sont présentées. Nous développons ensuite deux travaux. Le premier présente notre approche sparse PLS et compare les résultats obtenus avec la PLS originale sur plusieurs jeux de données. Deux variantes sont proposées, pour un contexte de soit de régression soit d'analyse canonique, suivant l'objectif de l'expérience biologique. Sur un des jeux de données, nous présentons une interprétation biologique détaillée des résultats obtenus. Le deuxième travail est un projet d'article comparant, au moyen de critères biologiques, les résultats de trois méthodes canoniques, afin de pouvoir guider le biologiste soucieux d'analyser son jeu de données avec une méthode adéquate.
   Nous concluons également cette deuxième partie par un bilan et des perspectives de travail futur.

3. Enfin la troisième partie présente deux contributions à des développements en biologie.
   (a) La première se place dans l'étude de la folliculogénèse chez la truie et a nécessité l'application d'approches connues, telles que le test de Fisher et Random Forests, dans le cadre de données déséquilibrées.
   (b) La deuxième contribution se place dans le cadre de l'élongation du placenta d'embryons bovin inséminés artificiellement. Cette étude s'est faite en trois temps :
   - l'identification de gènes prédicteurs de stade de développement lors de l'application de OFW+CART, de OFW+SVM, de *Random Forests* et du test de Fisher ;
   - la validation de l'expérience biologique (*microarray*) sur ces gènes sélectionnés grâce à une PCR qualitative ;
   -enfin, la validation du caractère prédictif du stade de développement embryonnaire, en mesurant l'expression de ces même gènes identifiés, mais sur de nouveaux embryons.

Première partie

Sélection de gènes pour des problèmes de classification supervisée

2. Contexte

2.1 Motivation

Les données transcriptomiques dont nous disposons consistent en l'expression de milliers de gènes ($p \simeq 1000 - 10000$) mesurés sur un nombre restreint de lames ou membranes ($n \simeq 50 - 100$). La finalité globale de ces expériences biologiques est de comprendre les interactions et régulations entre gènes présents sur les puces à ADN dans des conditions données. Plus précisément, dans le cas par exemple de données de cancer, l'analyse statistique peut répondre à trois types de questions (Dudoit *et al.*, 2002) :

- identifier de nouvelles classes de tumeur grâce à l'aide des profils d'expression des gènes (*classification*, *apprentissage non supervisé*) ;
- classer des individus dans des classes de cancer connues (*analyse discriminante*, *apprentissage supervisé*) ;
- identifier des gènes *marqueurs* caractérisant le ou les différents cancers (*sélection de variables*).

Nous nous sommes principalement intéressés à ce dernier point dans le cadre de la classification supervisée, à la fois sur des données publiques de cancer très répandues dans la littérature et sur des données non publiques issues de l'INRA (Bonnet *et al.*, 2008; Tosser-Klopp *et al.*, 2008 ; ainsi que section 12) ou de projets européens (EADGENE[1], de Koning *et al.*, 2007; Jaffrézic *et al.*, 2007; Sorensen *et al.*, 2007; Watson *et al.*, 2007). D'un point de vue biologique, la sélection de variables (ici les gènes) devrait permettre de développer des tests de diagnostic pour détecter la maladie et pourrait aussi apporter plus de connaissances sur les caractéristiques de telle ou telle tumeur dans le cas par exemple de données oncologiques.
D'un point de vue statistique (Guyon *et al.*, 2003), la sélection d'un sous-ensemble de variables pertinentes permettrait d'améliorer la performance de prédiction des méthodes de classification et ainsi de passer outre le fléau de la haute dimensionalité de ces données (*the curse of dimensionality*), d'accélérer le temps de calcul de ces méthodes et enfin de comprendre le processus sous-jacent ayant "généré" ces données grâce à la lecture des représentations graphiques suffisamment simples à interpréter.
L'analyse statistique des données transcriptomiques comporte plusieurs étapes importantes résumées dans la figure 2.1 (Allison *et al.*, 2006). Dans cette partie, nous nous sommes principalement concentrés sur la classification supervisée et la sélection de vari-

[1] European Animal Disease Genomics Network of Excellence for Animal Health and Food Safety, `http://eadgene.info/NewsandEvents/EADGENEEvents/DataAnalysisWorkshopNov2006/tabid/166/Default.aspx`

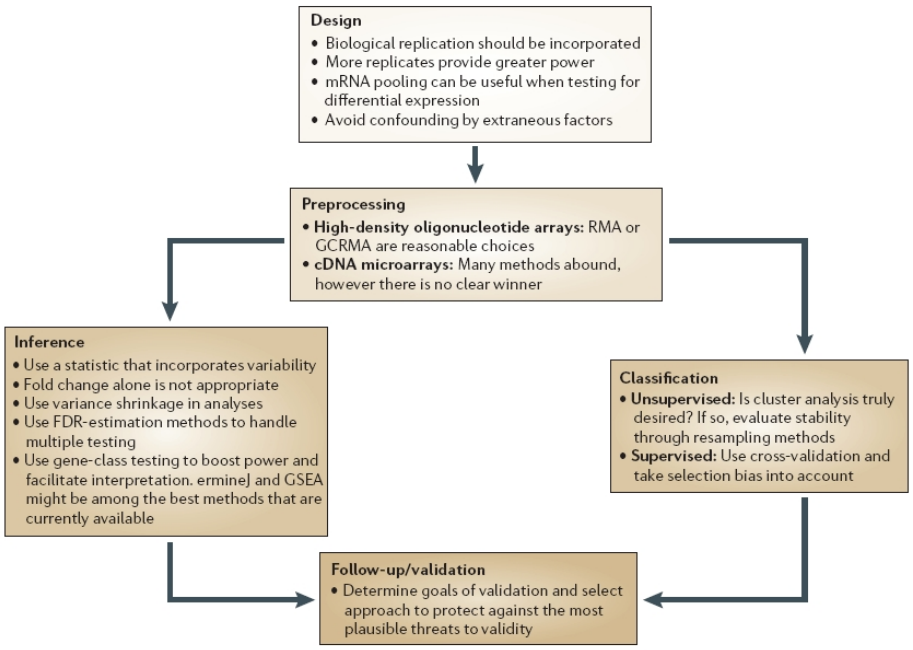


Figure 2 | **Guidelines for the statistical analysis of microarray experiments.** The flow chart indicates the guidelines for each relevant stage of a microarray study. FDR, false-discovery rate; RMA, robust multi-array average (includes a modification, CGRMA).

Fig. 2.1: Directives pour l'analyse statistique de données de transcriptome (source Allison *et al.*, 2006).

ables (les gènes), bien que les étapes de normalisation et d'inférence aient été prises en compte lors de nos analyses statistiques (Jaffrézic *et al.*, 2007; Bonnet *et al.*, 2008; Tosser-Klopp *et al.*, 2008 et section 12).

2.2 Classification supervisée et sélection de variables

2.2.1 Classification supervisée

La littérature concernant la sélection de gènes pour des données de transcriptome étant très vaste, nous nous intéressons dans cette partie uniquement aux méthodes dites de classification supervisée ainsi que leurs applications, laissant de côté d'autres méthodes couramment utilisées pour réduire la dimension telles que l'Analyse en Composantes Principales (Hastie *et al.*, 2000, *gene shaving*), la régression Partial Least Squares (PLS, Antoniadis *et al.*, 2003; Boulesteix, 2004 qui traitent la classification comme de la régression). Par ailleurs, nous nous focalisons uniquement sur l'extraction de variables à proprement parler et non pas sur la construction de nouvelles variables artificielles pour réduire la dimension.
Nous disposons de l'expression de p gènes mesurés sur n puces à ADN. Ces données sont

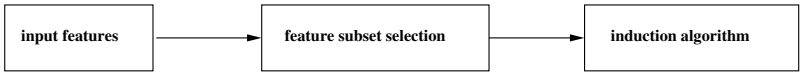


Fig. 2.2: Méthode filtre : la sélection se fait indépendamment du classifieur (source John *et al.*, 1994).

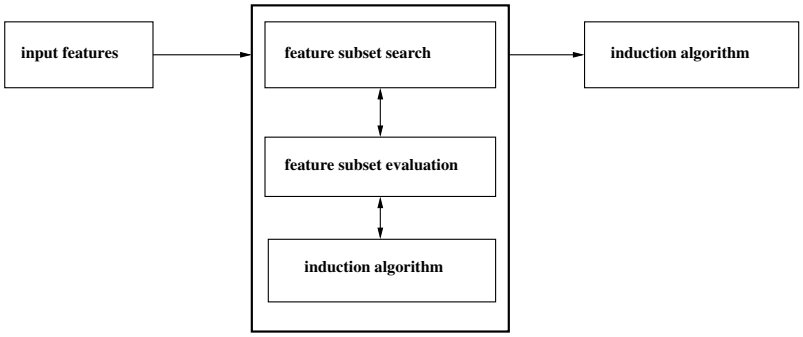


Fig. 2.3: Méthode de type *wrapper* : le classifieur est utilisé lors de la sélection (source John *et al.*, 1994).

stockées dans une matrice $X = x_{ij}$ de taille $n \times p$, où x_{ij} représente le niveau d'expression du gène j sur la puce i. Dans le cadre de la classification supervisée, chaque puce appartient à une classe biologique connue k, $k = 1, \ldots, K$ et l'on assignera à chaque observation (puce i) une modalité codée $y_i \in \{1, \ldots, K\}$ pour $i = 1, \ldots, n$. Nous nous plaçons dans le cas de données de grande dimension, où $p >> n$.

2.2.2 Sélection de variables

Dans la littérature du Machine Learning, trois classes de méthodes de classification et sélection de variables sont considérées et présentées dans les revues bibliographiques de Kohavi & John (1997); Blum & Langley (1997); Guyon *et al.* (2003) :

1. les méthodes *filtre*,
2. les méthodes de type *wrapper*,
3. les méthodes de type *embedded*.

Les méthodes filtre. Les méthodes filtre (figure 2.2) sont souvent considérées comme une étape de pre-processing pour sélectionner les gènes différentiellement exprimés. Elles consistent à tester chaque gène indépendamment des autres et à les ordonner selon un critère (par exemple une p-valeur). Des exemples basiques de méthodes filtres sont les tests de Student ou Fisher dans le cadre d'une ANOVA. Dans l'une des premières études comparatives de méthodes de classification, Dudoit *et al.* (2002) proposent de pré-filtrer les gènes sur la base du rapport de la somme des carrés inter et intra groupe :

$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k 1_{y_i=k}(\bar{x}_j^k - \bar{x}_j)^2}{\sum_i \sum_k 1_{y_i=k}(x_j^i - \bar{x}_j^k)^2}$$

où $\bar{x}_j$ est la valeur moyenne de l'expression du gène j sur toutes les observations et $\bar{x}_j^k$ la valeur moyenne de l'expression du gène j pour les observations de la classe k. Ils comparent ensuite les performances de certaines méthodes de classification telles que les k plus proches voisins (k-NN), Classification and Regression Trees (CART, Breiman *et al.*, 1984) et l'Analyse Linéaire Discriminante (LDA) sur une sélection de 30 à 50 gènes.
Le principal avantage des méthodes filtre est leur efficacité calculatoire et leur robustesse face au surapprentissage (ou surajustement). Malheureusement, ces méthodes ne tiennent pas compte des interactions entre gènes et tendent à sélectionner des variables comportant de l'information redondante plutôt que complémentaire (Guyon *et al.*, 2003). De plus, les sélections de gènes faites dans une étude préalable ne tiennent absolument pas compte de la performance des méthodes de classification appliquées dans la deuxième étape de l'analyse (Kohavi & John, 1997).

Les méthodes de type wrapper. Le concept des méthodes de type *wrapper* a été introduit par John *et al.* (1994). Ces méthodes consistent en l'évaluation de la performance de sous-ensembles de gènes de manière successive, prenant ainsi en compte les interactions entre variables. Ainsi, l'algorithme de sélection "entoure" (*wrapp*) la méthode de classification (encore appelé *classifieur*) qui évalue la performance (figure 2.3). La recherche d'un tel sous-ensemble de gènes optimal requiert certaines définitions au préalable (Guyon *et al.*, 2003) : comment rechercher dans l'espace des variables tous les sous-ensembles possibles, comment évaluer la performance de prédiction d'une méthode d'apprentissage pour guider la recherche, quand arrêter l'algorithme. Bien entendu, une recherche exhaustive est un problème NP-difficile et incalculable lorsque p est grand ; il nécessite des approximations des calculs d'optimisation. Le risque de surapprentissage est grand si le nombre d'observations n est insuffisant et le nombre de variables à sélectionner doit être choisi par l'utilisateur. Enfin, le plus grand désavantage de ces méthodes est le temps de calcul qui devient vite important dès que p est grand.
John *et al.* (1994) et Aha & Bankert (1995) furent les premiers à démontrer (de façon empirique) que la stratégie *wrapper* était supérieure à la stratégie filtre en terme de performance de classification.

Les méthodes de type embedded. Les méthodes *embedded* incorporent la sélection de variables lors du processus d'apprentissage, sans étape de validation, pour maximiser la qualité de l'ajustement et minimiser le nombre de variables. Un exemple très connu est celui de CART, où les variables sélectionnées sont celles présentes au niveau de la division de chaque noeud. D'autres approches incorporent des stratégies de recherche "gourmandes" (*greedy*) du type *sélection forward* ou *élimination backward* conduisant à des sous-ensembles de variables imbriqués. Lors d'une sélection forward, les variables sont progressivement incluses dans des sous-ensembles de plus en plus grands, alors que la méthode backward part de l'ensemble de variables initiales et élimine progressivement les variables les moins pertinentes. Selon la revue de Guyon *et al.* (2003), ces approches seraient bien plus avantageuses en terme de temps de calcul que les méthodes de type *wrapper* et seraient robustes face au problème de surajustement. La sélection forward

semble plus efficace en temps de calcul que l'élimination backward pour générer des sous-ensembles imbriqués de variables. Cependant, on risque de sélectionner des sous-ensembles de variables peu pertinents dans le premier cas, puisque l'importance des variables n'est pas évaluée en fonction des autres variables (qui ne sont elles pas encore incluses).

2.2.3 Agrégation de modèles

Certaines méthodes de classification sont extrêmement sensibles à de petites perturbations du jeu de données initial. C'est le cas par exemple de CART, dont la construction peut varier de façon notable suite à la modification de quelques valeurs. Si un modèle a une faible capacité de généralisation, c'est-à-dire si sa variance est grande, alors une des solutions proposées par Breiman (1996) est d'agréger les classifieurs. De ce fait, la variance du modèle est réduite mais l'interprétation du classifieur devient moins aisée (Kohavi & John, 1997).
Breiman (1996) proposa donc d'agréger des arbres CART pour réduire leur variance, chaque arbre étant estimé à partir d'un échantillon bootstrap. Il introduisit le concept de *bagging* (pour *bootstrap aggregating*), créant des échantillons d'apprentissage "perturbés" de la même taille que l'échantillon original, dans lequel chaque arbre est estimé sur un échantillon obtenu par tirage bootstrap. Plus tard il introduisit une variante sous la forme de forêts aléatoires (*Random Forests*, Breiman, 2001, cf. section 2.3.3) qui introduit une perturbation aléatoire supplémentaire dans la construction des noeuds des arbres afin de les rendre plus "indépendants" les uns des autres.
Une autre façon d'agréger des modèles est le *boosting* (Freund & Schapire, 1997), où les observations sont pondérées de façon adaptative et les classifieurs agrégés par vote. A chaque itération, la pondération d'une observation mal ajustée est renforcée tandis que l'agrégation finale prend en compte la qualité prédictive de chaque modèle. Dettling & Buhlmann (2003) ont en particulier appliqué l'approche *LogitBoost* (Freund & Schapire, 1997) avec des arbres de décision sur des données d'expression de gène, dans le cadre de classification de tumeurs. Leurs résultats conduisent à une meilleure prédiction avec du boosting plutôt qu'avec un arbre de classification simple. Par la suite, Dettling (2004) a présenté une approche hybride combinant à la fois bagging et boosting (*BagBoosting*) et montre que les résultats sont compétitifs avec d'autres méthodes, dont la méthode d'agrégation d'arbres Random Forest de Breiman (2001). Dans ce travail, nous n'avons pas abordé le cas du boosting.

2.3 Exemples de méthodes de sélection de variables

Dans cette section, nous allons détailler trois méthodes de sélection de variables, qui ont ensuite été comparées à l'approche que nous avons développée par ailleurs (Lê Cao *et al.*, 2007b,a, cf. sections 3 et 4). Les deux premières utilisent comme classifieurs binaires le *Support Vector Machines* (SVM), et la troisième le classifieur multiclasse CART.

Support Vector Machines. Rappelons brièvement les formulations des problèmes à résoudre dans les cas des SVM à marge douce (problème non séparable). Les deux classes sont codées 1 et -1 et on considère l'échantillon d'apprentissage $\{\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_i, \dots, \mathbf{x}_n\} \in \mathbb{R}^p$ et la modalité associée $\{y_1, y_2, \dots, y_i \dots y_n\} \in \{-1, +1\}$.
L'algorithme de support vecteur cherche le meilleur hyperplan qui sépare les données. On recherche donc le couple $(\mathbf{w}, b)$, $\mathbf{w} \in \mathbb{R}^p$, b réel, tel que :

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}_i + b \geq +1 - \xi_i \qquad & \text{pour} \qquad y_i = +1 \\ \mathbf{w} \cdot \mathbf{x}_i + b \leq -1 + \xi_i \qquad & \text{pour} \qquad y_i = -1 \\ & \text{sous la contrainte} \qquad \xi_i \geq 0 \quad \forall i. \end{aligned} \tag{2.1}$$

En combinant (2.1) on obtient :

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \tag{2.2}$$

où $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$, $b =< y_i - \mathbf{w} \cdot \mathbf{x}_i >$ est la valeur du *biais*, $\mathbf{w}$ est appelé le vecteur normal de l'hyperplan et $\alpha_i \geq 0$. Pour maximiser la marge qui est la distance du point le plus proche à l'hyperplan et qui est égale à $2/||\mathbf{w}||$, on minimise $\frac{1}{2}||\mathbf{w}||^2$, sous la contrainte (2.2). Le problème s'écrit alors sous forme quadratique :

$$\begin{aligned} & \min_{\mathbf{w},b} ||\mathbf{w}||^2 \\ \text{tel que} \quad & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \end{aligned} \tag{2.3}$$

et les points $\mathbf{w} \cdot \mathbf{x}_i + b = \pm(1 - \xi_i)$ sont les *vecteurs supports* qui définissent la solution trouvée. La formulation duale du lagrangien s'écrit finalement

$$\begin{aligned} & \min \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j - \sum_{i=1}^{m} \alpha_i \\ \text{sous les contraintes} \qquad & 0 \leq \alpha_i \leq C \qquad \text{et} \qquad \sum_i \alpha_i y_i = 0. \end{aligned} \tag{2.4}$$

C correspond à une pénalité pour les points mal séparés et les vecteurs supports sont les observations pour lesquelles $\alpha_i \neq 0$. Ainsi, le problème est résolu de façon à ce que le nombre de vecteurs supports (et donc le nombre de paramètres dans le modèle) soit petit. En contrôlant le nombre de vecteur supports, le SVM permet donc de contrôler le sur-apprentissage et de résoudre le problème (2.4). Le lecteur pourra se référer aux articles de Burges (1998); Vapnik (1999); Scholkopf & Smola (2001) pour une description plus détaillée des algorithmes.

2.3.1 Recursive Feature Elimination

RFE (Guyon *et al.*, 2002) est une méthode de type *embedded* basée sur l'élimination backward et utilisant les Support Vector Machines (SVM) pour sélectionner un sous-ensemble de gènes optimal non redondants. La méthode repose sur le fait que chaque

élément $\mathbf{w}_j$ du vecteur de poids $\mathbf{w}$ sur chaque variable j est une combinaison linéaire des observations et que la plupart des α_i sont nuls, exceptés pour les observations *support* dans le problème (2.4). Par conséquent la mesure $\mathbf{w}$ peut être directement reliée à l'importance des variables dans le modèle SVM. En effet la mesure $\mathbf{w}^2$ est une mesure de pouvoir prédictif. Ainsi, les variables j de plus petit poids $\mathbf{w}_j^2$ seront progressivement éliminées dans la procédure RFE.
L'algorithme RFE consiste en les étapes suivantes :

- Initialiser :
  $\mathbf{s} = [1, 2, \ldots, p]$ le sous-ensemble de variables courant ;
  $\mathbf{r} = \emptyset$ le sous-ensemble de variables sélectionnées et rangées par ordre d'importance.
- Tant que $\mathbf{s} \neq \emptyset$:
  1. apprentissage du SVM sur les données $\mathbf{x}_{i\mathbf{s}}$ et y_i ;
  2. calcul du vecteur de poids $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$;
  3. calcul du critère de rang $c_j = \mathbf{w}_j^2$ pour tout indice j dans $\mathbf{s}$;
  4. trouver la variable avec le plus petit critère de rang : $f = arg\min(c)$;
  5. mettre à jour $\mathbf{r} = [f, \mathbf{r}]$;
  6. eliminer la variable avec le plus petit critère de rang : $\mathbf{s} = [\mathbf{s} \setminus f]$.

Pour accélérer le temps de calcul, Guyon *et al.* (2002) proposent d'éliminer plusieurs variables à la fois, bien que la performance de classification puisse en être affectée. Dans ce cas-là, on obtient non pas un critère de rang sur des variables, mais un critère de rang sur des sous-ensembles de variables qui sont imbriqués les uns dans les autres. Il est conseillé d'éliminer des sous-ensembles de variables de tailles différentes, par exemple la moitié des variables dans $\mathbf{s}$, afin d'obtenir une densité d'information suffisamment importante sur les derniers gènes éliminés (et qui seront classés les premiers).
Si les variables sont éliminées une à une comme le propose l'algorithme initial, Guyon *et al.* (2002) nous mettent en garde sur la pertinence des variables du plus haut rang : seul le *sous-ensemble de variables* sélectionné est optimal, et non pas les variables de plus haut rang considérées individuellement. En effet RFE est une méthode de type *wrapper* qui va avoir tendance à sélectionner des variables comportant de l'information complémentaire, améliorant ainsi la tâche de classification. Les gènes considérés un à un dans la sélection ne contiennent que peu d'information pertinente.
Il est important de noter que RFE ne s'intéresse pas à la recherche du sous-ensemble de taille optimale, mais donne une mesure d'importance sur chaque variable ou groupe de variables. Les auteurs ne spécifient pas précisément comment choisir les différentes tailles des sous-ensembles récursivement éliminés dans la deuxième approche. Il semblerait que, suivant la complexité du jeu de données, il convienne d'appliquer l'approche originale, comme ils le font sur les données Colon de Alon *et al.*, 1999, plutôt que la deuxième approche (données Leukemia de Golub *et al.*, 1999).
RFE a été appliqué pour l'analyse de données de transcriptome (Ramaswamy *et al.*, 2001) et a suscité le développement de nombreuses variantes. SVM-RFE-annealing est basé sur un algorithme de recuit simulé, de façon à éliminer un nombre important de variables lors des premières itérations, puis à réduire le nombre de variables

éliminées. Ceci permet de réduire de façon significative le temps de calcul. Cette méthode, très proche de RFE, nécessite de fixer (par l'utilisateur) le nombre de variables à sélectionner. Un autre exemple, est celui de SVM-RCE (Yousef *et al.*, 2007) pour Recursive Cluster Elimination, afin de sélectionner des ensembles de gènes corrélés. La critique principale de Yousef *et al.* (2007) concernant RFE est que certains gènes avec des petits poids (et donc jugés peu informatifs) peuvent être importants mais leur rang bas est le résultat de la présence d'autre gènes dominants hautement correlés à ces derniers. Ceci soulève le problème de gènes correlés et comportant de l'information redondante. Devraient-ils être tous présents dans la sélection, au risque de reclasser avec un rang plus bas des gènes comportant une autre information complémentaire ? Ou de nécessiter une sélection de taille plus grande ? Cette question ne semble pas avoir été posée aux utilisateurs directement concernés, les biologistes.
D'autres variantes ont aussi été proposées et le lecteur pourra se référer à Tang *et al.* (2007), Mundra & Rajapakse (2007) ou Zhou & Tuck, 2007 (MSVM-RFE pour le cas multiclasse).

2.3.2 l_0 norm SVM

La méthode l_0 norm SVM de Weston *et al.* (2003) est une méthode de type *embedded* encore basée sur le classifieur SVM. Elle consiste à la fois à minimiser le taux d'erreur d'apprentissage et à sélectionner des variables en une étape unique.
Les auteurs partent d'une formule générale du problème quadratique (2.3) :

$$\min_{\mathbf{w}} ||\mathbf{w}||_k^k$$

sous la contrainte

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i$$

où $||\mathbf{w}||_k = (\sum_{j=1}^{p} |\mathbf{w}_j|^k)^{1/k}$ est la l_k norme de $\mathbf{w}$. Si $k = 2$, on retombe sur le problème de SVM à marge optimale (dans le cas séparable).
Dans le cas où $k \to 0$, Weston *et al.* (2003) définissent ce qui est couramment appelée la "norme" zéro de $\mathbf{w}$, bien que cela ne soit pas une norme au sens mathématique, de la façon suivante : $||\mathbf{w}||_0^0 = card\{\mathbf{w}_j | \mathbf{w}_j \neq 0\}$, où $card\{\mathbf{w}_j | \mathbf{w}_j \neq 0\}$ représente le nombre de variables sélectionnnées.
La régularisation appliquée sur w en minimisant la norme l_0 permet donc de répondre directement au problème de sélection de variables (puisque certains poids seront nuls), tout en améliorant le pouvoir discriminant du classifieur en une seule étape d'optimisation (Weston *et al.*, 2003). Le problème d'optimisation à résoudre est donc le suivant :

$$\begin{aligned} &\min_{\mathbf{w}} ||\mathbf{w}||_k^k \\ \text{sous les contraintes} \quad &y_i(\mathbf{w}.\mathbf{x}_i + b) \geq 1 - \xi_i \quad \text{et} \quad ||\mathbf{w}||_0^0 \leq r \end{aligned} \tag{2.5}$$

où r est le nombre de variables sélectionnées fixé par l'utilisateur.
Le problème (2.5) est cependant NP-difficile et nécessite de nombreuses approximations que nous ne présenterons pas ici. Finalement l'algorithme l_0 norm consiste à modifier le SVM de la façon suivante :

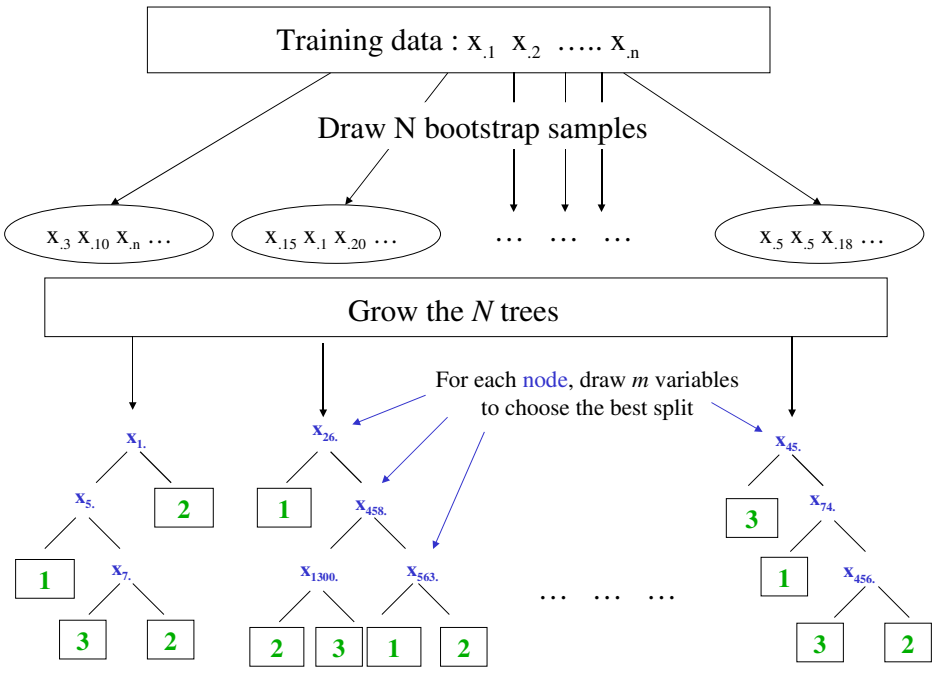


Fig. 2.4: Construction des forêts aléatoires avec l'inclusion de deux aléas, pour un problème multiclasse (3 classes, indiquées en vert).

1. construire un SVM linéaire (avec régularisation l_1 ou l_2) ;
2. standardiser les observations en les multipliant par la valeur absolue du vecteur $\mathbf{w}$ obtenu ;
3. répéter l'étape 2 jusqu'à convergence.

Cette méthode rappelle RFE basée sur l'élimination des variables de plus petits poids $|\mathbf{w}_j|$. Encore une fois les réserves exprimées par Yousef *et al.* (2007) peuvent aussi s'appliquer ici. Les résultats obtenus par Weston *et al.* (2003) sur des données de transcriptome (Colon, Alon *et al.*, 1999 ; Lymphoma, Alizadeh *et al.*, 2000) présentent des taux d'erreur de classification (sur échantillon test) similaires à RFE, bien que le temps de calcul soit un peu plus important que RFE. Pourtant, cette approche intéressante au niveau statistique, n'a pas suscité autant d'engouement au niveau applicatif que RFE. Une des raisons problables est que l'article ne donne aucune évaluation biologique des gènes sélectionnés (la méthode est uniquement validée sur ses résultats statistiques).

2.3.3 Les forêts aléatoires

Random forests ou forêts aléatoires (Breiman, 2001) est une méthode de type *wrapper* devenue très populaire pour la classification et la sélection de variables, dans diverses applications (Izmirlian, 2004; Strobl *et al.*, 2007; Bureau *et al.*, 2005; Diaz-Uriarte & Alvarez de Andres, 2006 pour des données biologiques, Svetnik *et al.*, 2003

pour données QSAR, Prasad *et al.* (2006) pour des données écologiques etc.). Cette approche qui semblait au départ très empirique (Breiman, 2001) fait maintenant l'objet de quelques études théoriques (consistance d'agrégation d'arbres dans Biau *et al.*, 2007). Le paradoxe surprenant des forêts aléatoires est que cette méthode arrive à tirer profit de la grande instabilité des classifieurs CART en les agrégeant. Cette approche combine deux sortes d'aléa qui améliore grandement la précision de prédiction : le bagging et une sélection aléatoire des variables lors de la construction des noeuds des arbres. Il en résulte à la fois un biais et une variance faibles du modèle. Chaque arbre (de classification ou de régression) est construit selon l'algorithme suivant (figure 2.4) :

1. tirer N échantillons boostrap $\{B_1, \ldots, B_N\}$ sur le jeu de données initial ; chaque échantillon B_k $(k = 1 \ldots N)$ sera l'échantillon d'apprentissage pour construire un arbre T_k non élagué ;
2. soit p le nombre total de variables dont on dispose ; à chaque noeud de l'arbre, m variables sont tirées aléatoirement $(m << p)$ pour déterminer la meilleure division de chaque noeud, m étant fixé au départ.

Les prédictions des N arbres sont ensuite agregées pour prédire la classe d'une nouvelle observation, par vote majoritaire pour la classification, ou en moyennant pour la régression. Les forêts aléatoires évitent de faire de la validation croisée sur un échantillon test pour estimer l'erreur de prédiction de la forêt. En effet, lors de la construction, une estimation interne de l'erreur de généralisation est calculée ainsi :

1. lors de la construction de chaque arbre T_k, environ 1/3 des observations ne sont pas tirées dans l'échantillon boostrap B_k et ne serviront donc pas à la construction de l'arbre ; ces observations appelées "Out-Of-Bag" (OOB) seront utilisées comme échantillon test interne pour chaque arbre ;
2. les prédictions OOB sont ensuite agregées et le taux d'erreur OOB estimé est calculé pour la forêt entière.

Ce taux d'erreur, qui semble en pratique précis et non biaisé, ne pourra en aucun cas être utilisé pour évaluer la performance d'une sélection de variables. En effet, Svetnik *et al.* (2003) montrent que l'estimateur de l'erreur OOB tend à faire du surapprentissage puisque l'évaluation de l'erreur ne se fait pas sur un échantillon test externe. L'évaluation de la performance d'une sélection ne peut donc se faire que sur un exchantillon test externe. Nous reviendrons sur le biais de l'évaluation de sélection de variables dans la section 2.5.1.
Le choix du nombre de variables m tirées aléatoirement pour partitionner chaque noeud peut être fixé par défaut à $\sqrt{p}$ (Liaw & Wiener, 2002; Svetnik *et al.*, 2003). En revanche, le nombre d'arbres N doit être fixé par l'utilisateur. Pour obtenir des résultats stables, en particulier si le nombre d'observations est petit, notre expérience a montré que le choix de N de l'ordre de 8000 à 15000 était approprié (Bonnet *et al.* (2008); Tosser-Klopp *et al.* (2008); Sorensen *et al.* (2007) ; section 12).

Deux mesures d'importance sont proposées de façon interne dans les forêts aléatoires pour faire de la sélection de variables.

– *Mean Decrease Accuracy.* Les observations OOB sont utilisées pour estimer l'importance des variables en évaluant leur contribution à la précision de la prédiction. Les valeurs de chaque variable dans les observations OOB sont aléatoirement permutées et l'on compare les classes prédites de ces nouvelles observations OOB par rapport à leur vraie classe. On moyenne ensuite cette erreur de prédiction pour tous les arbres.

– *Mean Decrease Gini.* Le critère d'hétérogénéité de Gini pour chaque noeud t de l'arbre est défini par $i(t) = 1 - \sum_k \hat{p}_k^2$ où $\hat{p}_k$ est la proportion relative d'observations appartenant à la classe k dans le noeud t. La qualité de la division s d'un noeud t est la décroissance $\Delta(s,t) = i(t) - p_R i(t_R) - p_L i(t_L)$ où p_R (p_L) est la proportion d'observations classées dans le noeud descendant droit (gauche). Ainsi l'importance de chaque variable est évaluée en calculant la décroissance totale résultant de la division du noeud sur cette variable, moyennée sur l'ensemble des arbres.

Attention ces mesures peuvent conduire à des résultats totalement différents si le jeu de données comporte peu d'observations et si certaines des classes ont des caractérisques (biologiques) très similaires (travail de DEA, Lê Cao, 2005).
Dans l'étude sur la folliculogénèse, la mesure *Mean Decrease Accuracy* s'est avérée très instable comparée à *Mean Decrease Gini*. En revanche, cette dernière peut être sujette au surapprentissage (communication personnelle avec Andy Liaw). En effet *Mean Decrease Gini* dépend de la construction interne (noeud à noeud) de chaque arbre, alors que *Mean Decrease Accuracy* dépend de la construction de chaque arbre, mais dans sa globalité. De manière générale, *Mean Decrease Accuracy* se place dans un contexte prédictif, alors que *Mean Decrease Gini* s'inscrit dans un contexte descriptif ou explicatif du modèle. Nous avons donc choisi d'utiliser la mesure *Mean Decrease Gini* pour l'étude biologique de Bonnet *et al.* (2008), et nous avons tenté de réduire l'instabilité des résultats en agrégeant plusieurs forêts aléatoires.
Diaz-Uriarte & Alvarez de Andres (2006) ont récemment présenté une méthode de type *embedded* basée sur l'élimination backward et les forêts aléatoires afin d'identifier un sous-ensemble optimal de gènes. La sélection se fait en éliminant progressivement les variables ayant une mesure *Mean Decrease Accuracy* faible, de façon à ce que l'erreur OOB estimée soit minimale. Attention dans ce cas particulier, l'estimation de l'erreur OOB est biaisée (trop optimiste) et ne sert pas à estimer l'erreur de généralisation (Ambroise & McLachlan, 2002).
Cette approche, finalement très similaire à RFE puisque seul le classifieur change, est très coûteuse en temps de calcul (temps de calcul d'une forêt aléatoire >> temps de calcul d'un SVM). Par ailleurs, l'instabilité de la mesure d'importance *Mean Decrease Accuracy* que nous avons évoquée peut faire douter de la pertinence des variables sélectionnées (on risque de trouver des variables sélectionnées très différentes si on relance plusieurs fois cette approche sur le même jeu de donnée).

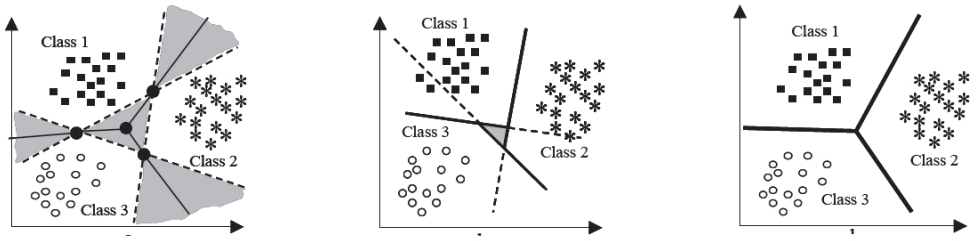


Fig. 2.5: Plusieurs façon de traiter le problème multiclasse avec les SVM : combiner des SVM binaires *1 vs. 1* (gauche) ou *1 vs. rest* (milieu) ou bien implémenter un SVM multiclasse (droite). Source Statnikov *et al.*, 2005.

2.4 Le cas multiclasse

2.4.1 Subdiviser en problèmes binaires

Bien que Guyon *et al.* (2003) soutiennent le fait que le cas multiclasse soit plus aisé pour la sélection de variables que le cas binaire (plus le nombre de classes est grand, plus on est sûr de tomber sur un sous-ensemble de variables non aléatoire pour une bonne classification), le problème est en pratique bien plus difficile à résoudre. En effet, la grande dimensionalité étant toujours présente, on se retrouve ici face à un nombre d'observations par classe limité (dû principalement à une question de coût expérimental), ce qui rend la précision de la prédiction de plus en plus mauvaise lorsque le nombre de classes augmente (Li *et al.*, 2004a). Ramaswamy *et al.* (2001); Yeang *et al.* (2001) furent parmi les premiers à s'intéresser aux problèmes multiclasses en choisissant de combiner les sorties de classifieurs binaires pour faire de la prédiction multiclasse, soit par vote pondéré, soit en utilisant k-NN ou les SVM binaires. Yeang *et al.* (2001) en particulier, notent que la plupart des erreurs de classification sont dues à des observations appartenant à des classes très similaires, plutôt qu'à des observations "aberrantes".
Certaines méthodes de classification binaires s'étendent naturellement au cas multiclasse. C'est le cas par exemple de l'analyse discriminante linéaire ou bien de CART. D'autres nécessitent de décomposer le problème multiclasse en plusieurs problèmes binaires de type une classe contre une autre (*1 vs. 1*) ou une classe contre le reste (*1 vs. rest*). Une autre solution est de définir des fonctions objectifs multiclasses (voir figure 2.5) . En particulier, la question a été soulevée de nombreuses fois avec les SVM. Weston & Watkins (1999) et Lee & Lee (2003), par exemple, ont proposé la résolution du problème d'optimisation quadratique pour du multiclasse directement dans le SVM plutôt que d'agréger des SVM binaires. Ces auteurs concluent que le nombre de supports est inférieur si l'on résout le problème multiclasse directement plutôt que d'agréger des SVM binaires. Néanmoins d'après Lee & Lee (2003), il reste bien moins coûteux de résoudre plusieurs petits problèmes binaires qu'un seul gros multiclasse. Des techniques de décomposition de calculs proposées par Hsu & Lin (2002) permettraient cependant d'améliorer de façon significative la vitesse de résolution de ces problèmes quadratiques. Le problème de diviser un problème multiclasse en plusieurs problèmes binaires pose

aussi la question du choix de la méthode d'agrégation : vote majoritaire, *least square estimation based weighting* (consiste à pondérer chaque SVM), *double layer hierarchical combining* (consiste à agréger les résultats des SVM dans un autre SVM), proposés par Kim *et al.* (2002). Cela pose aussi la question du type de classifieur binaire : le SVM *1 vs. rest* peut donner de mauvais résultats si plusieurs classes sont similaires (Lee & Lee, 2003) ; en revanche le SVM *1 vs. 1* peut comporter une grande variabilité, puisque chaque classifieur binaire est calculé sur un sous-ensemble d'observations très restreint et ne permet qu'une structure de coût unique en cas de mauvaise classification, plutôt que différents coûts possibles. Ce problème est en partie lié au problème de classes déséquilibrées que nous abordons plus tard (cf. section 2.4.2).
Une comparaison de certaines approches SVM multiclasses (approche de Weston & Watkins, 1999; Lee & Lee, 2003, *1 vs. rest* et *1 vs. 1*) a été présentée dans Statnikov *et al.* (2005) dans le cas de données de transcriptome avec pré-sélection de gènes via une méthode filtre. Notons aussi l'approche récemment présentée par Zhang *et al.* (2008) qui proposent des SVM multiclasse avec une technique de régularisation adaptative, de manière à pondérer les variables importantes et donc de permettre la sélection de variables (application sur les données SRBCT de Khan *et al.*, 2001).
De manière générale, le problème de classification multiclasse reste un problème ouvert.

2.4.2 Le problème des classes déséquilibrées

S'il est commun d'obtenir un nombre de classes supérieur à deux dans les données de transcriptome, il est tout aussi commun de faire face à des classes déséquilibrées. La principale raison est qu'en général, la classe d'interêt est la plus rare, ce qui rend l'obtention des données difficile. Le problème des classes déséquilibrées a été très peu soulevé jusqu'à présent : Lee *et al.* (2003), dans le cas de modèles linéaires, Chen *et al.* (2004), et plus récemment Eitrich *et al.* (2007) et Qiao & Liu (2008) dans le cadre de la classification. Le principal danger, dans le cadre de la classification et de la sélection de variables, est qu'un classifieur a pour but de minimiser le plus possible le taux d'erreur de classification et donc le taux d'erreur de la classe majoritaire, au détriment de la classe minoritaire. Cette approche a notamment des répercussions importantes lors de la sélection de variables, puisque les gènes sélectionnés discrimineront en priorité les classes majoritaires qui ne sont pas forcément les plus pertinentes biologiquement.
Pour les forêts aléatoires, Chen *et al.* (2004) ont proposé deux approches différentes pour équilibrer au mieux les classes et introduire une pénalité plus forte lors d'une mauvaise classification de la classe minoritaire. La première approche, *Balanced Random Forests* (BRF) est basée sur la technique du ré-échantillonnage. Chaque arbre est construit sur le même nombre d'observations dans les classes majoritaires et minoritaires (tirage avec remise). La deuxième approche, *Weighted Random Forests* (WRF) est basée sur le coût d'apprentissage (*cost sensitive learning*). Des poids sont introduits dans l'algorithme de RF, en premier lieu dans la construction de l'arbre, où des poids sur les classes sont utilisés dans le calcul du critère de Gini pour diviser les noeuds et en deuxième lieu lors de l'assignation de la classe du noeud terminal.
BRF risque fortement le surapprentissage si le nombre d'observations des classes mi-

noritaire est très faible car cette approche de *down sampling* utilise finalement peu les observations présentes dans les classes majoritaires. En revanche, WRF semble bien adapté à nos données et l'inclusion de poids de façon interne dans l'algorithme est une approche intéressante dont nous nous sommes inspirés par la suite (cf. Lê Cao *et al.*, 2007a et section 4).
Dans le cadre des SVM, Qiao & Liu (2008) ont proposé d'inclure, dans la formulation du problème d'optimisation quadratique du SVM multiclasse de Lee & Lee (2003), une procédure d'apprentissage de poids adaptatifs (*adaptive weighted learning*), de manière à pondérer de façon optimale chaque classe.
De manière générale, le double problème de sélection de variables *et* de prise en compte des classes minoritaires dans un shéma déséquilibré a été peu traité dans la littérature (Eitrich *et al.*, 2007).

2.5 Evaluation des approches statistiques

2.5.1 Evaluation de la performance

L'évaluation de la performance des méthodes de classification et sélection de variables reste très difficile compte tenu du nombre restreint d'observations. Comme le souligne Dudoit *et al.* (2002), plus d'observations seraient nécessaires pour obtenir un taux d'erreur raisonnablement précis. Il est en effet souvent impossible d'avoir recours à un échantillon test externe et la validation de la performance doit souvent être calculée sur le jeu de donnée d'apprentissage.
Par ailleurs, dans le cas de la sélection de variables, de nombreux auteurs (Ambroise & McLachlan, 2002; Reunanen *et al.*, 2003; Svetnik *et al.*, 2003; Simon *et al.*, 2003; Allison *et al.*, 2006) nous mettent en garde contre le problème du biais dans la sélection. En effet de nombreux articles ont présenté des résultats extrêmement optimistes car l'estimation du taux d'erreur se faisait sur l'échantillon d'apprentissage. Dans le cas où n est petit et ne permet pas d'obtenir un "vrai" échantillon test, la performance de la sélection devrait être évaluée de la façon suivante :

1. diviser son jeu de données initial en B échantillons d'apprentissage et B échantillons "test" (par bootstrap ou validation croisée) ;
2. pour chaque échantillon b, $b = 1, \ldots, B$:
   (a) apprendre la sélection de variables sur l'échantillon d'apprentissage ;
   (b) tester sur l'échantillon "test" la performance de la sélection ; dans le cas boostrap, tester aussi sur l'échantillon d'apprentissage (erreur *in bag*).
3. Agréger ensuite les erreurs de classification obtenues dans chaque phase de test, soit en moyennant lorsqu'on l'on fait de la validation croisée (CV), soit en pondérant les erreurs *in bag* et *out bag* comme proposé par la méthode *e.632+* de Efron & Tibshirani (1997).

En revanche, notons que la sélection finale de variables (celle que l'on présentera au biologiste) se fait sur le jeu de données initial.

Comme n est petit, cette estimation de l'erreur de généralisation ne devrait être considérée que comme élément de comparaison face à d'autres approches concurrentes, et non pas comme une estimation précise et réelle de l'effet de la sélection de gènes. Par conséquent, cette erreur estimée ne peut en aucun cas guider la taille optimale de la sélection.
Il est difficile de savoir quelle méthode d'estimation du taux d'erreur serait la moins biaisée possible. Braga-Neto & Dougherty (2004) proposent d'utiliser l'estimateur *.632* boostrap ou bien l'estimateur de resubstitution *in bag* dans ces cas extrêmes. De leur côté Fu *et al.* (2005) proposent une méthode combinant à la fois validation croisée et bootstrap.
Un consensus semble néanmoins se dégager (Ambroise & McLachlan, 2002, Diaz-Uriarte & Alvarez de Andres, 2006, Statnikov *et al.*, 2005 etc.) pour le choix de la méthode e.632+ d'Efron & Tibshirani (1997) et le 10-CV. Cependant, lorsque le nombre d'observations par classe est inférieur à 10, ce qui est souvent le cas pour les données de transcriptome (non publiques), toutes ces techniques d'estimation de l'erreur de généralisation restent très limitées, voire non applicables.

2.5.2 Valider la pertinence des résultats

Les premières analyses de données de transcriptome furent présentées dans la fin des années 90 (par exemple Golub *et al.*, 1999; Alon *et al.*, 1999; Alizadeh *et al.*, 2000). Les méthodes statistiques utilisées étaient des méthodes filtres pour sélectionner des gènes différentiellement exprimés, suivies de méthodes de classification (Eisen *et al.*, 1998) ou les cartes topologiques de Kohonen (2001) (*Self Organising Map*, SOM). La validation des résultats se faisaient alors essentiellement sur la pertinence biologique des listes de gènes sélectionnés.
Des méthodes issues du Machine Learning furent ensuite introduites. Ces approches, plus sophistiquées, ont essentiellement été appliquées par des statisticiens. Par conséquent peu de publications proposent à la fois une expertise statistique et biologique. On pourra citer cependant les articles de Guyon *et al.* (2002) pour RFE et Brown *et al.* (2000) présentant ces deux aspects dans le cas d'application de SVM.
Un très grand nombre d'approches assez complexes ont ensuite été développées pour le cas de la sélection de variables en grande dimension, et plus particulièrement pour les données de transcriptome. Cependant, ces méthodes ne sont que rarement validées sur des jeux de données réels. Dans la plupart des cas, le recours aux données simulées est largement utilisé, bien que plusieurs auteurs reconnaissent que la tâche est complexe (Yeung & Burmgarner, 2003; Nguyen & Rocke, 2004). Il nous semble en effet impossible de simuler de façon réaliste des données aussi bruitées et avec des structures de variance/covariance d'une telle complexité. De plus, les données simulées sont souvent issues d'un modèle mathématique sous-jacent. Par conséquent, lors d'une étude comparative de différentes méthodes statistiques, la méthode dont les hypothèses de départ sont similaires à celles des données simulées sera favorisée. Nykter *et al.* (2006), en proposant une simulation "réaliste" de ce type de données, montrent la réelle complexité de la tâche. Ainsi, des études comparatives récentes semblent de moins en moins faire

appel à des données simulées pour valider les approches statistiques (Li *et al.*, 2004b, Pirooznia *et al.*, 2008, Zeng *et al.*, 2008).
Une autre façon courante de valider les résultats d'une nouvelle méthode statistique est de l'appliquer sur des jeux de données publiques et de comparer la performance obtenue à celles d'autres algorithmes. Un exemple poussé à l'extrême est celui de Lee *et al.* (2005) qui ont comparé 21 méthodes de classification et 3 méthodes de sélection de gènes (méthodes filtre) sur 7 jeux de données.
Ces travaux utiles pour la communauté "Data Mining" ne s'intéressent cependant pas à l'interprétation biologique des résultats et donc à la pertinence des sélections de gènes obtenues. Par ailleurs, comme mentionné dans la partie 2.5.1, il arrive souvent que le critère de validation statistique ne soit pas applicable pour de vrais jeux de données fournis par le biologiste. Enfin, le nombre de nouvelles méthodes developpées (souvent très sophistiquées) est tel qu'il devient maintenant difficile pour le biologiste de choisir une approche plutôt qu'une autre si la pertinence biologique des résultats n'est pas soulevée lors de ces applications.
C'est pourquoi nous avons jugé qu'il était absolument indispensable de se placer à l'interface entre la statistique et la biologie. Lors de nos travaux, nous avons ainsi systématiquement choisi de nous appuyer sur la validité des résultats au niveau biologique en plus de comparaisons quantitatives lorsque celles-ci étaient possibles.

2.6 Plan de la partie

Dans cette première partie, nous présentons trois travaux basés sur le développement et l'application d'un algorithme de type *wrapper* pour la sélection de gènes, dans le cadre de données de transcriptome.
L'articles méthodologique et l'article appliqué ont tous deux bénéficié d'un apport important concernant l'interprétation biologique des résultats. La méthode développée a ensuite fait l'objet d'une implémentation sous forme d'un package `R`[2], maintenant accessible sur le CRAN. Le troisième article présente ce package `R`.
Nous proposons pour terminer quelques éléments de discussion et perspectives concernant cette première partie.

[2] The Comprehensive R Archive Network, `http://cran.r-project.org/`

3. Article méthodologique

Cet article présente une extension du méta-algorithme de sélection de variables "*Optimal Feature Weighting*" (OFW) de Gadat & Younes (2007) dans le cadre des données de transcriptome, avec les classifieurs CART et SVM. Plusieurs approches sont comparées : une méthode filtre basique (test de Student) et des méthodes de types *wrapper* ou *embedded* utilisant les classifieurs SVM (RFE, l_0 norm SVM) ou bien CART (Random Forests), sur trois jeux de données de cancer connus. Les comparaisons des différentes approches ont été faites sur des critères statistiques, mais aussi biologiques.

L'originalité de ce travail tient à l'algorithme proposé ainsi qu'à la place importante que nous avons voulu donner à l'interprétation biologique des différentes listes de gènes sélectionnés. En effet, compte tenu du petit nombre d'échantillons, il nous est apparu primordial de pouvoir évaluer les approches sur des critères pragmatiques qui soient parlants pour le biologiste. Nous montrons la complémentarité des approches filtre-*wrapper* et soulignons le fait qu'il n'existe pas une seule et unique méthode pouvant répondre à un problème biologique souvent complexe.

Cet article a été publié dans la revue Statistical Application in Genetics and Molecular Biology (Vol. 6 : Iss. 1, Article 29, 2007).

Selection of biologically relevant genes with a wrapper stochastic algorithm

Kim-Anh Lê Cao,[1,2] Olivier Gonçalves,[3] Philippe Besse [1] and Sébastien Gadat [1]

Abstract

We investigate an important issue of a meta-algorithm for selecting variables in the framework of microarray data. This wrapper method starts from any classification algorithm and weights each variable (*i.e.* gene) relatively to its efficiency for classification. An optimization procedure is then infered which exhibits important genes for the studied biological process.
Theory and application with the SVM classifier were presented in Gadat and Younes (2007) and we extend this method with CART. The classification error rates are computed on three famous public databases (Leukemia, Colon and Prostate) and compared with those from other wrapper methods (RFE, l_0 norm SVM, Random Forests). This allows to assess the statistical relevance of the proposed algorithm. Furthermore, a biological interpretation with the Ingenuity Pathway Analysis software outputs clearly shows that the gene selections from the different wrapper methods raise very relevant biological information, compared to a classical filter gene selection with T-test.

Introduction

Performing a feature selection algorithm has several important applications in the field of microarray data analysis. First, to determine which genes contribute the most for the biological outcome (*e.g.* cancerous *vs.* normal cells) and in which way they interact to determine this outcome. Second, to predict the outcome when a new observation is presented. It is unlikely that thousands of genes do explain the class membership of a microarray and it is hence wise to use a dimensional reduction technique. This also provides practical aspects with machine learning methods: it avoids the "curse of dimensionality" that leads to overfitting when the number of variables is too large.

Features can generally be selected with two different approaches: either explicitly (filter methods) or implicitly (wrapper methods). The aim of the filter methods is

[1]Institut de Mathématiques, Université de Toulouse et CNRS (UMR 5219), F-31062 Toulouse, France
[2]Station d'Amélioration Génétique des Animaux UR 631, Institut National de la Recherche Agronomique, F-31326 Castanet, France
[3]Laboratoire de Biologie des Protistes, UMR CNRS 6023, Université Blaise Pascal, 63177 Clermont-Ferrand, France

to measure the relevance of each gene. Variables are usually ordered with statistical tests and microarrays are classified with the few good-ranked selected variables. In this case, note that the selection is totally independent from the classification method (Dudoit et al., 2000 and Golub et al., 1999). The main advantages are robustness against overfitting and low cost computation, but these methods may fail to select the most "useful" features and usually disregard the interactions between the features. On the other hand, wrapper methods measure the usefulness of a set of features by exploring the subsets space. This search can be performed either with heuristic or stochastic techniques (*e.g.* simulated annealing, genetic algorithms). These methods find the "useful" variables, but are prone to overfit. Moreover, when dealing with numerous variables, an exhaustive subspace search is computationally untractable. They generally yield greedy and costly algorithms since each iteration consists in selecting smaller and smaller subsets of variables (Guyon et al., 2001, Diaz-Uriarte and Alvarez de Andrés, 2006).

These latter wrapper methods have been successfully applied on several benchmarks but suffer from lack of mathematical justification. Furthermore, they are all dedicated to one special baseline classifier that is used for constructing the decision rule. Gadat and Younes (2007) proposed a wrapper approach which does not depend on the classifier and can numerically quantify the efficiency of each gene. It uses stochastic approximations that still cover a large portion of the search space to avoid local minima. This reaches to subset selections of discriminative genes that hence hold useful information on the microarray experiment.

The two main objectives of this paper are first to numerically compare the performances of different wrapper methods by estimating the classification error rate with the e.632+ bootstrap method (Efron and Tibshirani, 1997) and second, to provide a comparison of the different gene selections based on their biological relevance. Note that we do not intend to optimize the size of the gene subset. We rather focus on the biological interpretation of the 50 first selected genes.

The optimal feature weighting procedure (ofw) from Gadat and Younes (2007) was initially applied with the classifier Support Vector Machines (SVM: Vapnik, 2000). We investigate the application of another classifier, Classification and Regression Trees (CART) on public microarray data sets (*Leukemia*: Golub et al., 1999, *Colon*: Alon et al., 1999 and *Prostate*: Singh, D. et al., 2002). We compare the results from these two wrapper methods ofw+SVM or ofw+CART to those obtained with other well known procedures: Recursive Feature Elimination (RFE: Guyon et al., 2002), Random Forests, (RF: Breiman, 2001) and l_0 norm SVM (l_0: Weston et al., 2003), as well as the widely used T-statistics. The classification error rates are displayed for each public data set and the biological relevancies of the gene selections are discussed with the Ingenuity Pathways Analysis software.

1 Method

We introduce the optimal feature weighting meta-algorithm (ofw) from Gadat and Younes (2007) that treats several classification problems with a feature selection task. In this section we explain the main theoretical derivations that are necessary to fully understand the algorithm and its application.

1.1 Optimal Feature Weighting Model

The particularity of this algorithm is that it does not depend on the classification procedure $\mathbb{A}$ used for classification. We consider a large set of genes $\mathcal{G}$ of size N expressed on two biological conditions (or classes) $\{\mathcal{C}_1, \mathcal{C}_2\}$. $\mathcal{G}$ can be either the total number of genes spotted on the microarray or a rather large gene subset. These N genes describe a signal $\mathcal{I}$. The optimization of any given classification algorithm $\mathbb{A}$ (*e.g.* SVM, CART, Nearest Neighbors ...) is explored by passing through $\mathbb{A}$ different subspaces of genes to improve its performance with time.

System energy

Let us define a positive weight parameter $\mathbb{P}$ on each of the genes in $\mathcal{G}$. After a normalization step, we can consider $\mathbb{P}$ as a discrete probability on the N genes. The goal is to learn a probability that fits the efficiency of each gene for the classification of $\mathcal{I}$ in $\{\mathcal{C}_1, \mathcal{C}_2\}$, so that important weights are given to genes with high discriminative power and lower weights to those that have a poor influence on the classification task.

Denote p any small integer compared to N (*e.g.* $p = 2\% * N$), a gene subset of size p has to be extracted from $\mathcal{G}$. Next definition properly establishes how to measure the goodness of $\mathbb{P}$ for the set of genes $\mathcal{G}$ and the two classes $\{\mathcal{C}_1, \mathcal{C}_2\}$.

Definition of system energy:
Given a probability $\mathbb{P}$ on $\mathcal{G}$ and $\epsilon(\omega)$ the measure of classification efficiency with any p-uple $\omega \subset \mathcal{G}^p$, the energy of the system at the point $\mathbb{P}$ is defined as the mean classification performance when ω is drawn with respect to $\mathbb{P}^{\otimes p}$ (with replacement) in $\mathcal{G}^p$, that is:

$$\mathcal{E}(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[\epsilon] = \sum_{\omega \subset \mathcal{G}^p} \mathbb{P}(w)\epsilon(w) \tag{1}$$

Note here that the energy $\mathcal{E}$ depends on the way we measure the classification efficiency on ω, denoted $\epsilon(\omega)$ all along this paper. Given any standard classification algorithm $\mathbb{A}$, $\epsilon(\omega)$ will be the error of $\mathbb{A}$ computed on the training set using the set of extracted features ω. For instance, if $\mathbb{A}$ is a SVM with a linear kernel, $\epsilon(\omega)$ will be the classification error of a linear SVM using only genes in ω to describe the signal in the training set.

The computation of the sum (1) is untractable since one cannot enumerate all subsets ω of $\mathcal{G}^p$, but we will provide a stochastic algorithm to optimize $\mathcal{E}$ in next section.

REMARK The more $\mathbb{P}$ enables to hold a discriminative gene g for classification (important weight on g and $\epsilon(\omega)$ small each time ω contains this gene g), the less $\mathcal{E}$. Minimizing $\mathcal{E}$ with respect to $\mathbb{P}$ will thus permit to exhibit the most weighted and thus the most discriminative genes. Hence, a natural measure of variable importance ranking will be read on the weight distribution $\mathbb{P}^\star$ minimizing $\mathcal{E}$.

1.2 Stochastic optimization method

This part provides an efficient way to minimize the energy $\mathcal{E}$ with a stochastic version of the standard gradient descent technique.

Remark first that the function $\mathcal{E}$ has to be minimized up to the constraints defined by a discrete probability measure on $\mathcal{G}$. Thus, the most natural way to optimize (1) is to use a gradient descent of $\mathcal{E}$ projected on the set of constraints. This leads to the next definitions.

DEFINITIONS:
We define the set $\mathcal{S}$ as the simplex of probability map on $\mathcal{G}$. We also denote by $\Pi_{\mathcal{S}}$ the affine projection of any point of $\mathbb{R}^N$ on the simplex $\mathcal{S}$. This natural projection $\Pi_{\mathcal{S}}$ of any point x can be computed in a finite number of steps as mentioned in Gadat and Younes (2007).
The Euclidean gradient of $\mathcal{E}$ is:

$$\forall g \in \mathcal{G} \qquad \nabla\mathcal{E}(\mathbb{P})(g) = \sum_{\omega \subset \mathcal{G}^p} \frac{C(\omega, g)\mathbb{P}(\omega)}{\mathbb{P}(g)}\epsilon(\omega) \tag{2}$$

where $C(\omega, g)$ is the number of occurrences of g in ω. The iterative procedure to update $\mathbb{P}$ is then given by

$$\mathbb{P}_{t+dt} = \mathbb{P}_t - \nabla\mathbb{P}_t dt \tag{3}$$

Of course (2) is numerically impossible to calculate, as one cannot enumerate all possible ω in $\mathcal{G}^p$ and a stochastic approximation is needed: the Euclidian gradient expression (2) can actually be seen as an expectation. Then, to deal with such gradient, a computable Robbins-Monro algorithm can be used, which gets similar asymptotic behavior as (3) (see for instance Gadat and Younes (2007), Kushner and Clark (1978)). With this stochastic method, the updated formula of $\mathbb{P}_n$ becomes:

$$\mathbb{P}_{n+1} = \Pi_{\mathcal{S}}\left[\mathbb{P}_n - \alpha_n \frac{C(\omega_n, .)\epsilon(\omega_n)}{\mathbb{P}_n(.)}\right] \tag{4}$$

where ω_n is any set of p genes sampled with respect to $\mathbb{P}_n$, and $\alpha_n = K/(n+1)$ for any positive constant $K > 0$ is the step of the algorithm. Note that the last expression

is always defined since when $\mathbb{P}_n(g) = 0$, we cannot draw this gene in ω_n and $C(\omega_n, g)$ vanishes.

Under mild conditions on the energy $\mathcal{E}$, one can show that this stochastic approximation algorithm converges to a critical point of $\mathcal{E}$. One can also prove the asymptotic normality result:

$$\frac{\mathbb{P}_n - \mathbb{P}_\infty}{\sqrt{\alpha_n}} \to \mathcal{N}(0, V),$$

where the covariance matrix V depends on the energy function $\mathcal{E}$. Further details can be found in (Benveniste et al., 1990).

1.3 Detailed algorithm

Let $\mathcal{G} = (\delta_1 \dots \delta_{|\mathcal{G}|})$, $\mu \in \mathbb{N}^*$ and η the stopping criterion.

- For $n = 0$ define $\mathbb{P}_0$ as the uniform distribution on $\mathcal{G}$
- While $|\mathbb{P}_{(n+\mu)} - \mathbb{P}_n|_\infty > \eta$:
  - extract ω_n from $\mathcal{G}^p$ with respect to $\mathbb{P}_{n,p} = \mathbb{P}_n^{\otimes p}$
  - construct $\mathbb{A}_{\omega_n}$ and compute $\epsilon(\omega_n)$
  - compute the drift vector $d_n = C(\omega_n, \cdot)\epsilon(\omega_n)/\mathbb{P}_n(\cdot)$
  - update $\mathbb{P}_{n+1} = \Pi_{\mathcal{S}}[\mathbb{P}_n - \alpha_n d_n]$
  - $n = n + 1$

2 Application

We first provide a short description of the two supervised algorithms we apply ofw to: Support Vector Machines (SVM) and Classification And Regression Trees (CART). We next shortly describe other feature selection methods that we compare to our approach.

2.1 Two baseline classifiers are applied to ofw

Support Vector Machines

SVM (Vapnik, 2000) is a binary classifier that attempts to separate the microarrays into $\mathcal{C}_1$ and $\mathcal{C}_2$ by defining an optimal hyperplane between the 2 classes up to a consistency criterion. Linear kernel SVMs are used here because of their good generalization ability compared to more complex kernels.

Classification And Regression Trees

CART (Breiman et al., 1984) is a multi-category classifier that is constructed through a recursive partitioning routine. It builds a classification rule to predict the class label of the microarrays based on the feature information following the Gini criterion. To avoid overfitting, trees are then pruned using a cross validation procedure. Note that CART is naturally unstable: a slight change in the features can lead to a very different construction of the tree.

2.2 Comparisons with existing ranking methods

We briefly present here the several algorithms we performed to compare our OFW approach with. Each of these methods follows the classical framework of feature selection algorithm. A training set is used to compute the rank (or relevancy) of each feature (or gene) and the error of the obtained gene selection is then computed on a test set. Thus, the input of each of these algorithms is simply the training set in our case.

Recursive Feature Elimination

RFE (Guyon et al., 2002) is a feature selection technique exclusively dedicated to SVM. It consists in computing a ranking criterion for all features using the SVM previously computed. Genes with the smallest ranking criterion are then recursively removed (with more than one feature per step for speed reasons). The idea is to construct several stacked feature subsets $\mathcal{G}_m \subset \mathcal{G}_{m-1} \subset \ldots \subset \mathcal{G}_1 = \mathcal{G}$ and find $\mathcal{G}_m$ that is optimal (on the basis of error rates metrics) and that leads to the largest margin of class separation. In this paper we will only focus on the gene ranks that are output from this method and not on the optimal size of the subset so as to compare the different methods. Indeed, all the presented methods do not necessarily give a stopping criterion for an optimal selection size.

l_0 norm SVM

Weston et al. (2003) proposed to minimize the l_0 norm of the normal vector from SVM to provide a way of selecting features and to minimize the training error in one step. As the problem is NP-hard, an approximation of the l_0 norm is proposed. This feature selection method has rarely been used yet in the context of microarray.

Random Forests

RF is a CART aggregation technique. The idea of Breiman (2001) was to introduce two sources of randomness. First with bagging: each unpruned tree is constructed on a bootstrap sample. Second, for each partition building step of the tree, the best variable is chosen among a fixed number of randomly selected variables. Trees are

then aggregated by majority vote. There is also an internal importance measure of the variables given by the forest that determines which predictors (*i.e* genes) are the most discriminative. Here we choose the "Mean Decrease Accuracy" measure that consists for each tree in randomly permuting the genes values that are not in the bootstrap sample (called "Out-Of-Bag" data) and computing the resulting classification error rate.

Diaz-Uriarte and Alvarez de Andrés (2006) proposed a backward feature selection procedure using RF that has not been applied here as the selection is often extremely small with no redundant genes.

Univariate filter method

One of the aim of this paper is to compare the gene selection using T-statistics to the ones resulting from the multivariate classification methods that were presented above. Note that the False Discovery Rate that controls the number of false positive genes was not applied here as we are selecting a fixed number of genes.

2.3 Public microarray data sets

We present the results obtained on three well known public data sets. *Leukemia* (Golub et al., 1999) compares two different types of leukemia (Acute Myeloid and Acute Lymphoplastic, ALL *vs.* AML) with 3860 genes and 72 microarrays. *Colon* (Alon et al., 1999) was obtained from cancerous or normal colon tissues with 2000 genes and 62 microarrays and *Prostate* (Singh, D. et al., 2002) also compared normal *vs.* cancerous prostate tissues with 102 microarrays and 12600 genes. These data sets will be refered as Leukemia, Colon and Prostate along this paper. We assumed the data sets correctly normalized.

2.4 Error rate assessment

We compared the error rates of all methods on each data set with the e.632+ bootstrap error estimate from (Efron and Tibshirani, 1997) that is adequate for small sample size data sets (Ambroise and MacLachlan, 2002) . The e.632 estimator is defined as $e.632 = .368R + .632B$ where R is the resubstitution error rate and B the ouf-of-bag bootstrap error rate. When the number of genes is much larger than the number of samples, the prediction rule usually overfits (R often equal 0). Efron and Tibshirani proposed the e.632+ estimate

$$e.632+ = (1-w)R + wB$$

with $w = \frac{.632}{1-.368r}$, $\quad r = \frac{B-R}{min(B,\gamma)-R}$, $\quad \gamma = \sum_{i=1}^{2} p_i(1-q_i)$ where r is an overfitting rate and γ the no-information error rate, p_i the proportion of samples of class $\mathcal{C}_i$, q_i the

proportion of samples assigned to class $\mathcal{C}_i$ with the prediction rule and $i = 1, 2$.

Note that e.632+ does not dictate the optimal number of features to select. The error rate estimates that are computed with respect to the number of selected features are only a way to compare the performances of the different methods. Remark at last that each algorithm needs to be learned on each bootstrap sample of the e.632+ bootstrap method to avoid any selection biais (Ambroise and MacLachlan, 2002). Concerning the performance assessment of a T-test selection we used a linear SVM as classifier. We assumed that although SVM is unrelated with this univariate method, it is well appropriate for this two-class problem.

2.5 Computing the efficiency of classification ϵ

The theoretical part showed that the ofw algorithm can be run with any classifier. However, computing the classification efficiency depends on the classifier. For ofw+CART, because of the unstable nature of CART, one needs to aggregate trees as in Breiman (1996) to reduce their variability. For iteration n, we launched B trees on B bootstrap samples on different ω_n^b drawn with respect to $\mathbb{P}_n$, where $b = 1, \ldots, B$. We then defined ϵ as the mean classification error rate on the out-of-bag samples.

No aggregation was needed with SVM, that is known to be very stable, and hence for this case B=1.

2.6 Computational amendements

For ofw+CART a mean gradient was computed that improved the speed of the algorithm

$$G_n = \frac{\sum_{i=1}^{n} \alpha_i \bar{d}_i}{\sum_{i=1}^{n} \alpha_i} \quad \text{with} \quad \bar{d}_i = \sum_{b=1}^{B} \frac{C(\omega_i^b, .)\epsilon(\omega_i^b)}{\mathbb{P}_i(.)}$$

where $\alpha_i = K/(i+1)$, $i = 1..n$ for any positive constant $K > 0$, as defined in equation (4).

Furthermore, to accelerate the computations, the data set Prostate that had a very high classification difficulty was filtered with a very large cut-off T-test p-value (we kept the genes below the p-value 0.1, which corresponded to 3584 remaining genes). Here we made the assumption that most genes are noisy or uninformative and can be

Colon / Prostate	ofwSVM	RFE	l_0	ofwCART	RF	T-test
ofwSVM	#	14	13	6	0	5
RFE	24	#	27	0	0	0
l_0	21	39	#	1	0	0
ofwCART	4	4	4	#	16	16
RF	6	5	4	17	#	36
T-test	7	5	3	12	31	#

Table 1: Number of genes shared by the several feature selection algorithms on Colon (upper triangle) and Prostate (lower triangle) for a selection of 50 genes.

Leukemia	ofwSVM	RFE	l_0	ofwCART	RF	T-test
ofwSVM	#	16	18	12	15	14
RFE		#	27	10	12	13
l_0			#	8	11	12
ofwCART				#	33	25
RF					#	32
T-test						#

Table 2: Number of genes shared by the several feature selection algorithms on Leukemia for a selection of 50 genes.

removed without affecting the biological study. Indeed, only a very small subset of genes do explain the outcome.

3 Results and discussion

3.1 Numerical results

3.1.1 Comparison of several selections

Table 1 displays the number of shared genes with the different methods when selecting 50 genes on the benchmarks Colon (upper triangular table) and Prostate (lower triangular table).

It first underlined the fact that all gene selections depended on the performed method as there were very few genes that were shared among all methods (less than 36 in Colon and 39 in Prostate). Furthermore, as expected, the methods could be divided in three groups: group 1 and 2 used either the classifier SVM (ofw+SVM, RFE and l_0) or CART (ofw+CART and RF) and group 3 is composed of the method T-test on its own.

Methods in the same group shared an important number of selected genes (for instance at least 13 genes in group 1 and 16 genes in group 2 for Colon). Conversely, the number of genes shared in-between groups was very low (0 to 6 between groups 1 and 2 for Colon). Compared with group 3, more than half of the genes selected with

RF were differentially expressed (meaning significant with the T-test) as well as about one third for the genes selected with ofw+CART.

The group 1 did not select many differentially expressed genes (0 to 5 for Colon). The difference is that SVM looks for non redundant genes which lead to a linear separation between the classes C_1 and C_2. These genes are not necessarily differentially expressed. On the other hand, when CART is constructed, it searches genes with the largest difference mean between the two classes. It was hence not surprising to find many differentially expressed genes in group 2.

These latter methods also selected discriminative subsets that were different from the T-test selection. The reason is that groups 1 and 2 take into account interactions between variables, as opposed to filter methods like T-test. The differences between these three groups are less striking in Table 2 on Leukemia as this data set seems more easy for the classification task (see section below). Nevertheless, we can observe that RF and ofwCART shared numerous genes that were also selected with T-Test.

3.1.2 Comparison of the error rate with selection

Figure 1 displays the e.632+ bootstrap error rates obtained with the different methods on the three data sets with respect to the number of selected genes. These graphs first showed the level of classification difficulty of the data sets: for all methods and for a number of selected genes going from 20 to 50, the e.632+ error rates varies from 1 to 6 % on Leukemia (**a**), from 10 to 30% on Colon (**b**), and from 5 to 23% on Prostate (**c**). This variation is even more accentuated as the methods do not have the same performance (Colon, Prostate). Leukemia got similar error rates for all methods as it is known to be relatively easy to classify.

The graphs showed that RF was the most stable and outperformed the other methods, except on Leukemia where it performs the worst. This can be explained as the forest is constructed only on the most discriminative variables and is less affected by noisy variables. Hence e.632+ or any error rate computation might not be appropriate to evaluate the performance of RF.

The T-test was not the most efficient as this univariate procedure eliminates noisy genes but does not yield compact non-redundant genes sets. Consequently, genes that are complementary but do not separate the data well are missed.

On the other hand, our two methods were competitive on the more complex data sets Colon and Prostate. On these data sets, ofw+CART gave better performance as CART searches for a non linear separation between features, which a linear SVM cannot perform. These graphs generally showed that a gene selection gives statistical good results when the size of the selection is large enough (greater than 10 genes, depending on the method) but not too large as noisy variables might then enter the selection. It is actually well known that it is impossible to achieve an errorless separation with a single gene. Better results are obtained with a combination of several genes. Note that

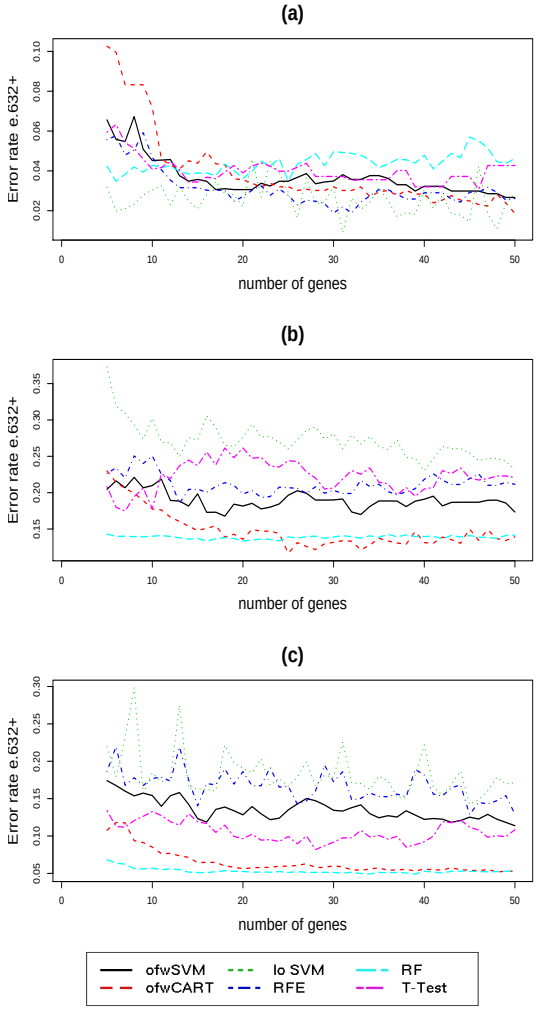


Figure 1: e.632+bootstrap error of several algorithms with respect to the number of genes on Leukemia (**a**) Colon (**b**) and Prostate (**c**).

Method / Criterion	T-test	RF	ofw CART	ofw SVM	l_0 SVM	RFE
Number of networks	7	4	4	6	3	3
Cancer term frequency in networks	3	2	1	0	1	1
Hematological disease term frequency in networks	0	0	0	0	1	1
Rank of the ontological term in the function list:						
Cancer	9	19	7	17	18	19
Hematological Disease	1	17	9	2	2	13
Number of surface markers	6	20	15	17	9	10
Number of genes associated with the ontological term:						
Leukemia	5	15	15	4	6	4
Myeloid Leukemia	0	3	4	2	0	0
Myeloid leukemia gene name		CD33 SPI1 TOP2B	TOP2B ITGB2 SPI1 CD33	TOP2B ITGB2		
Genes involved in the signaling pathways:						
NfKappaB		TRA@		TRA@	NFKBIA	TRA@
IL4	STAT6			BLVRB		
IL6			IL8			HSPB1
Wnt/Beta Catenin	GNAQ	TCF3	TCF3			
JAK/STAT	STAT6					

Table 3: Analysis of gene selections resulting from several feature selection algorithms on Leukemia data set. Comparisons of gene lists through IPA outputs with several criteria assessing global or specific information.

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 PPP2R5		PPP2RC
p38/MAPK					IL1R2, MEF2	MEF2
Wnt/Beta Catenin	CDH3	CDH3	CSNK2A2	CDH3		PPP2R5C

Table 4: Analysis of gene selections resulting from several feature selection algorithms on Colon data set. Comparisons of gene lists through IPA outputs with several criteria assessing global or specific information.

we did not determine here one optimal gene subset. Only the biological interpretation will give some clue about the relevance of the different selections.

3.2 Biological interpretation and discussion

Bioanalysis strategy

In order to elaborate an accurate assessment of the biological relevancy of the various tested methods, we analyzed all lists of 50 selected genes through Ingenuity Pathways Analysis [1] (IPA). IPA was chosen for two main reasons, first for its accuracy: IPA Ontology presents 25 times more classes than Gene Ontology (GO) and 85 high level functions compared to 3 for GO; and second because it supplies a more objective performance estimation compared to manual curating. Hence, with this strategy, we

[1] Ingenuity®Systems, www.ingenuity.com

Criterion \ Method	T-test	RF	ofw CART	ofw SVM	l_0 SVM	RFE
Number of networks	6	8	7	7	9	14
Cancer term frequency in networks	3	4	1	3	3	6
Renal and Urological disease term frequency in networks	0	0	0	0	0	0
Rank of the ontological term the in function list:						
Cancer	10	14	6	5	1	2
Renal and Urological Disease	49	0	33	59	0	0
Lipid Metabolism	25	28	27	36	17	15
Number of genes associated with the ontological term:						
Cancer	17	13	12	13	9	13
Prostate Cancer	4	3	3	1	1	4
Prostate cancer gene name	HPN SAT NME1 TGFB3	HPN TGFB3 GSTP1	FOLH1 HPN CLU	FOXO1A	SERPINB5	COX5A HOXC6 PMAIP1 SERPINB5
Genes involved in the signaling pathways:						
C21 steroid hormon metabolism				HSD11B1		
Androgen and Estrogen metabolism				HSD11B1		
Estrogen receptor signaling					CDK7	
Fatty acid metabolism					CYP4F2	CYP4F2
Wnt/Beta Catenin	TGFB3 NME1	TGFB3		WIF1 TLE4	PPP2R5E TLE4	WIF1 PPP2R5E
Pyrimidine or Purine metabolism	GUCY1A3 ATPGV1G1 DPYSL2		NME1	AOX1	RRM1	RRM1
PI3K/AKT				FOXO1A	PPP2R5E	
ERK/MAPK					PPP2R5E PAK1	PPP2R5E PAK1

Table 5: Analysis of gene selections resulting from several feature selection algorithms on Prostate data set. Comparisons of gene lists through IPA outputs with several criteria assessing global or specific information.

will focus more on global functions associated with a list of genes (*integrative biology*) than on a gene function associated with one gene only. This might allow to identify relevant genes present in a canonical pathway that were not selected with any statistical method.

We explored three outputs from IPA to generate performance indicators of a selected gene list: the *networks* that identify the interactions between the genes, and the most significant *functions* and *signaling pathways* generated by this gene list. This significancy is measured with a p-value of Fisher's exact test determining the probability that each biological function and disease assigned to a gene network or to a gene list was due to chance only. Concerning the canonical pathways, this significancy is furthermore measured by a ratio of the number of genes that map to a given pathway divided by the total number of genes that map to the canonical pathway generated by the gene list. More documentation about IPA can be found online.

The subsequent procedure was followed. First, we uploaded gene identifiers into the IPA application. Each gene identifier was mapped to its corresponding gene object in the Ingenuity Pathways Knowledge Base (IPKB). These genes, called "Focus Genes", were overlaid onto a global molecular network developed from information contained in the IPKB. Networks of these Focus Genes were then algorithmically generated based on their connectivity. Next, the functional analysis of a gene network identified the biological functions and diseases that were the most significant for those given genes. We also took into account the ranks of the most relevant biological functions and the canonical pathways were also considered.

An important remark In this interpretation we do not propose new information for cancer cause as the molecular data set depends entirely on the experimental setting that

was chosen by the biologists. The aim of this section is simply to check if our statistical results are not biologically aberrant and therefore contain relevant information that would need further experimental proof. The relevant selected genes can be called "predictive" from a statistical point of view (as they are selected on the basis of their predictive power), but from a biological point of view we do not pretend that these genes predict a cancer. The statisticians do hope that the selected genes might be predictive but the biologists can only evaluate the informative characteristics of these genes.

Leukemia data set

The aim of this data set was to select molecular markers distinguishing two leukemia variants arising from lymphoid precursors (Acute Lymphoblastic Leukemia, ALL) or from myeloid precursors (Acute Myeloid Leukemia, AML). Table 3 displays the biological performance estimated for each gene selection method.

In order to check the information quality from a selected gene list, several parameters were defined with various accuracy degrees. The potential abundance of information was first related to the number of networks. The more numerous the generated networks, the more varied the suggested biological clues. One gene list could also be considered as biologically relevant if ontological terms such as "Cancer" or "Hematological Disease" were linked to the networks of interacting genes or found well positionned, according to the p-values functions. We also focused on general leukemia molecular markers and more specifically on AML or ALL markers (Carroll et al., 2006, Pui et al., 2004), as well as surface marker gene families. These latter encode cell surface proteins that would be useful in distinguishing lymphoid from myeloid lineage cells as it was previously demonstrated for the CD33 gene (Drexler, 1987, Malask et al., 2006).

Canonical pathways did not reveal enough relevant differences between the gene lists to compare the methods. Networks generated by IPA were more numerous for the gene list selected by the filter method. It suggests that this method chooses less biologically interconnected genes compared to the wrapper methods. This could be explained by the fact that filter methods disregard the interactions between the features.

When looking for ontological terms, representative of leukemia pathology ("Cancer" and "hematological Disease"), no clear difference arose from any method as they were all well ranked in IPA interacting gene networks or function lists. Surface gene markers found in the networks were mostly selected with ofw+CART, RF and ofw+SVM, suggesting particular biological relevancy for those selected gene lists. Genes linked with "Leukemia" term or more precisely with "Myeloid Leukemia" terms were mostly selected in the lists given by the wrapper methods.

Compared to the wrapper methods, the filter method selected a poor number of general molecular markers linked to the leukemia pathology and surface markers distin-

guishing AML from ALL. No gene that was linked to Myeloid Leukemia was selected. On the other hand, wrapper methods gave very complementary and relevant gene lists. Three particular methods, ofw+CART, RF and ofw+SVM selected genes that are known to be involved in leukemia pathology *i.e.* TOP2B, ITGB2, SPI1, CD33). This trend was confirmed when we manually curated the gene lists proposed by all methods. ofw+CART, RF and ofw+SVM selected a set of genes involved at different biological level of the leukemia pathology (see for instance this non exhaustive list: CD33, ZYX, CCND3, TOP2B, SPI1, ITGB2, CCNI, NFIC, KPNB1).

To sumarize, we found that the T-test gene selection brought very general cancer-related information and much less information directly related to leukemia pathology than the CART or SVM based wrapper methods. The CART based methods proposed candidates that are linked to Myeloid Leukemia.

Colon data set

The objective of this data set was to select genes distinguishing tumor from normal sample. This is a particularly challenging problem since initial composition of the two types of cells are very different. Indeed, the high composition of tumor richness in epithelial cells and normal tissues in smooth muscle cells produce an important biological parameter that biases cancer-related genes tracking for tumor *vs.* normal cells (Guyon et al., 2002). Biological relevancy of the gene selections was assessed in the same manner as for the Leukemia data set with networks and function lists evaluation (Table 4). We chose ontological terms specific to colon cancer pathology such as "Cancer" and "Gastrointestinal Disease". Ontological terms linked to initial cells composition were also exploited to explain the performances of all methods ("Tissue Morphology", "Skeletal and Muscular System Development and Function"). Specific genes of colon cancer were also taken into account as well as specific signaling pathways.

For this particular data set, the number of networks generated by IPA for any gene selection method was similar. Principal differences arose from ontological functions of those networks. Indeed, rich cancer-related networks were generated from CART-classifiers methods as opposed to poor ones coming from the other methods. For any gene selection, the rank of the ontological term "Gastrointestinal Disease" in the function list was surprisingly low (line 5 of Table 4). An explanation of this particular feature could lie under the biological sample composition which is very rich (or too rich) in smooth muscle cells for normal tissue or in epithelial cells for tumor tissue. Interestingly, the ontological terms in IPA function lists "Skeletal and Muscular System Development and Function" or "Tissue Morphology" terms were always on top, lowering the rank of the "Gastrointestinal Disease" term. This observation was also reinforced by the larger number of genes linked with those last functions, comparing to those linked with the "Cancer" term. Therefore, exploitation of functions list analysis does not favor one method against another, as the sample biological composition bias

precludes straightforward detection of specific colorectal cancer genes.

When analyzing IPA canonical pathways, Wnt, MAPK and AKT signaling pathways that were (or supposed to be) involved in colorectal cancer, were mainly generated with gene selection involving SVM and CART methods (Oikonomou et al., 2006, Segditsas et al., 2006). Hence, SVM-classifiers (followed by CART-classifiers) seemed to select biologically relevant genes or signaling pathways, even in a data set that has a largely biased gene expression profile.

To summarize, we found that all methods selected genes that were more related to cell composition than to the pathology of interest. Very few colon cancer genes were identified. Despite the important biological biais, the wrapper methods were able to select complementary and relevant genes associated with relevant pathways.

Prostate data set

The identification of gene markers that might help to distinguish tumourous prostate from healthy prostate samples was the main purpose of this third data set. As for colorectal tumor, epithelial content of prostate tumor samples was significantly higher than in normal samples (79 *vs.* 27 %). This results in gene expressions correlated with epithelial content that may preclude cancer-related gene efficient tracking (Singh, D. et al.). Results are displayed in Table 5. We focused this time on the specific ontological terms "Cancer" and "Urological Disease". Prostate cancer specific genes and known deregulated signaling pathways were also used to determine more precisely the relevancy of the different selections.

All lists uploaded into IPA generated the same number of networks (except for RFE that was much higher) and were all linked with the ontological term "Cancer". This term was very well ranked for all function lists whereas, as observed for the Colon data set, the specific "Urological Disease" term was low ranked. When analysing ontological terms ranked in between, we noted a prevalence for functions involving cell proliferation, regulation of gene expression, lipid metabolism and nucleic acids, *i.e.* biological cell functions that are well known to be involved in prostate cancer disorders (Foley et al., 2004). The number of genes linked with ontological term "Cancer" was the same in any selection. When we focused on specific prostate cancer genes, all gene selection methods brought complementary information. For instance, CART-based methods selected HPN, a gene coding for a transmembrane serine protease involved in colony formation of prostate cancer cell lines (Dhanasekaran et al., 2001). SVM-based methods l_0 norm SVM and RFE selected SERPINB5, a gene coding for a serpin peptidase inhibitor involved in binding of prostate cancer cell lines Tahmatzopoulos et al. (2005).

IPA canonical pathways gave various information depending on the different selections. With the SVM-based methods, we observed signaling pathways involved in prostate cancer pathology (Terry et al., 2006) such as Wnt, MAPK, AKT, pyrimidine

and purine signaling pathway. In particular, l_0 norm SVM and RFE selections highlighted the Fatty Acid Metabolism and ofw+SVM the Androgen Signaling Pathway that is actually targeted for prostate cancer therapy (Singh, P. et al., 2002). Hence, SVM-based methods seemed to select here more relevant signaling pathways than the other methods.

To summarize, the T-test and the wrapper methods selected very complementary sets of genes related to prostate cancer in spite of the cell composition bias. All methods were also able to select complementary and relevant genes associated with relevant pathways.

4 Conclusion

The analysis of these three public data sets was performed at two levels. Statistically, we showed that the stochastic algorithm from Gadat and Younes could be applied to microarray data with two classifiers SVM and CART. ofw+CART and ofw+SVM gave excellent results compared to other well known wrapper methods. We also showed that the selected gene lists mostly depended on the chosen classifier.

Biologically, we showed that the relevancy did not only depend on the chosen method but also on the biological sample nature. Indeed, when applying these methods on a simple data set Leukemia, ofw+CART, RF and ofw+SVM proposed very relevant gene lists compared to the others. With a more complex biological matrix like in Colon or Prostate, the expression pattern are mixed between constitutive gene expression (*i.e.* expression of a large majority of genes involved in physiological characteristics of a tumor or normal cell) and cancer gene expression. In this setup, we rather observed a global complementarity of the biological information brought by the different selections. However, SVM-based methods seemed to propose interesting signaling pathways for Colon and Prostate data sets.

To summarize, we highlight the fact that the method statistically performing the best prediction does not necessarily give the most interesting biological results. In fact, the application of different methods on the same data set can highlight complementary relationships between the selected genes. Hence, to bring more information, one should not only consider the common features selected between the methods, but also the divergent ones. This means that there is not only one single method that answers a biological question: complementary approaches should be performed to analyze the data.

Availability

The code sources of ofw+SVM (in C++) and ofw+CART (in $\mathbf{R}$[2]) are available on the web site http://www.lsp.ups-tlse.fr/Biopuces/ofw/codesource/. An R package is currently being implemented but can be available upon request to the corresponding author.

References

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, **96**, 6745-6750.

Ambroise, C. and McLachlan, G. (2002) Selection Bias in Gene Extraction in Tumour Classification on Basis of Microarray Gene Expression Data. *Proc. Natl. Acad. Sci. USA*, **99**(10), 6562-6566.

Carroll W.L.,Bhojwani, D., Min, D.J., Moskowitz, N. and Raetz, E.A. (2006) Childhood Acute Lymphoblastic Leukemia in the Age of Genomics. *Pediatric Blood Cancer*, **46**(5), 570-578.

Benveniste, A., Métivier M. and Priouret, P. (1990) Adaptive Algorithms and Stochastic Approximations. (Springer-Verlag).

Breiman, L., Friedman, J.H., Olshen, R. A. and Stone, C.J. (1984) Classification and Regression Trees. (Wadsworth, Belmont, CA).

Breiman, L. (1996) Bagging predictors. *Machine Learning*, **24**, 123-140.

Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-22.

Dhanasekaran, S.M., Barrette, T.R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K.J., Rubin, M.A. and Chinnaiyan, A.M. (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature*, **412**(6849), 822-6.

Diaz-Uriarte, R. and Alvarez de Andrés, S. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**(3).

Drexler, H.G. (1987) Classification of acute myeloid leukemias : a comparison of FAB and immunophenotyping. *Leukemia*, **1**(1010), 697-705.

[2]The Comprehensive R Archive Network, http://cran.r-project.org/

Dudoit, S., Fridlyand, J. and Speed, T. (2000) Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *J. Am. Stat. Assoc.*, **97**, 77-87.

Efron, B. and Tibshirani R.J. (1997) Improvements on cross-validation: the e.632+ bootstrap method. *Journal of American Statistical Association*, **92**, 548-560.

Foley, R., Hollywood, D., and Lawler, M. (2004) Molecular pathology of prostate cancer: the key to identifying new biomarkers of disease. *Endocrine Related Cancer*, **11**(3), 477-488.

Gadat, S. and Younes, L. (2007) A Stochastic Algorithm for Feature Selection in Pattern Recognition. *Journal of Machine Learning Research*, **8**, 509-547.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

Guyon, I. and Weston, J. and Barnhill, S. and Vapnik, V. (2002) Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389-422.

Kushner, H. and Clark, D.S. (1978) Stochastic Approximation Method for Constrained and Unconstrained Systems. (Springer-Verlag).

Oikonomou, E. and Pintzas, A. (2006) Cancer genetics of sporadic colorectal cancer: BRAF and PI3KCA mutations, their impact on signaling and novel targeted therapies. *Anticancer Res.*, **26**(2A), 1077-84.

Maslak, P.G., Jurcic, J.G. and Scheinberg, D.A. (2006) Monoclonal antibodies for the treatment of acute myeloid leukemia. *Curr Pharm Biotechnol.*, **7**(5), 343-69.

Pui, C.H., Schrappe, M., Ribeiro, R.C. and Niemeyer, C.M. (2004) Childhood and adolescent lymphoid and myeloid leukemia. *Hematology Am Soc Hematol Educ Program*, **1**, 118-45.

Segditsas, S. and Tomlinson, I. (2006) Colorectal cancer and genetic alterations in the Wnt pathway. *Oncogene* **25**, 7531-7537.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander E.S., Loda M., Kantoff P.W., Golub T.R. and Sellers W.R. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **2**, 203-9.

Singh, P., Uzgare, A., Litvinov, I., Denmeade, S.R. and Isaacs, J.T. (2006) Combinatorial androgen receptor targeted therapy for prostate cancer. *Endocr Relat Cancer.*, **3**, 653-666.

Tahmatzopoulos, A., Sheng, S. and Kyprianou, N. (2005) Maspin sensitizes prostate cancer cells to doxazosin-induced apoptosis. *Oncogene*, **24**, 5375-5383.

Terry, S., Yang, X., Chen, M.W., Vacherot,F. and Buttyan, R.(2006) Multifaceted interaction between the androgen and Wnt signaling pathways and the implication for prostate cancer. *J Cell Biochem.*, **99**, 402-410.

Vapnik, V. (2000) The nature of statistical learning theory. (Springer-Verlag).

Weston, J., Elisseeff, A., Schölkopf, B. and Tipping, M. (2003) Use of the Zero-Norm with Linear Models and Kernels Methods. *Journal of Machine Learning Research*, **3**, 1439-1461.

4. Article appliqué

Cet article traite le cas particulier mais bien connu dans les données de transcriptome de la sélection de variables dans le cadre de données multiclasses déséquilibrées. En effet il est plus intéressant pour le biologiste d'obtenir des listes de gènes riches en information, plutôt que de l'information redondante portant sur la classe majoritaire, qui est souvent la plus évidente mais la moins pertinente biologiquement.
Dans ce travail, nous soulevons la complexité de tels jeux de données ainsi que la difficulté à trouver des critères statistiques d'évaluation adaptés. L'algorithme OFW est appliqué dans le cas multiclasse avec le SVM binaire *1 vs. 1* et CART, qui s'adapte naturellement au cas multiclasse. Les résultats de ces deux approches semblent donner des résultats compétitifs par rapport à la méthode filtre basique du test de Fisher et à Random Forests, bien que plus instables.
Un travail préliminaire d'interprétation biologique est inclus sur les données de folliculogènese (Bonnet *et al.*, 2008) et souligne l'importance de l'interprétation des résultats dans ce genre d'études comparatives.

Cet article a été soumis à la revue Computational Statistics and Data Analysis (octobre 2007), suite à une participation à la conférence IASC 07 - Statistics for Data Mining, Learning and Knowledge Extraction, Aveiro, Portugal, 2007. Cet article est actuellement en seconde lecture.

Multiclass classification and gene selection with a stochastic algorithm

Kim-Anh Lê Cao[1,2], Agnès Bonnet,[3] and Sébastien Gadat [1]

Abstract

Microarray technology allows for the monitoring of thousands of gene expressions in various biological conditions, but most of these genes are irrelevant for classifying these conditions. Feature selection is consequently needed to help reduce the dimension of the variable space. Starting from the application of the stochastic meta algorithm "Optimal Feature Weighting" (OFW) for selecting features in various classification problems, focus is made on the multiclass problem that wrapper methods rarely handle. From a computational point of view, one of the main difficulties comes from the commonly unbalanced classes situation when dealing with microarray data. From a theoretical point of view, very few methods have been developed to minimize any classification criterion, compared to the 2-class situation (*e.g.* SVM, l_0SVM, RFE...).
The OFW approach is developed to handle multiclass problems using CART and *one-vs-one* SVM as classifiers. The results are then compared with those obtained with other multiclass selection algorithm (Random Forests and the filter method F-test), on five public microarray data sets with various complexities. Statistical relevancy of the results is assessed by measuring and comparing the performances of these different approaches. The aim of this study is to heuristically evaluate which method would be the best to select genes classifying the minority classes. Application and biological interpretation are then given in the case of a pig folliculogenesis study.

Introduction

When dealing with microarray data, one of the most important issues to improve the classification task is to perform feature selection. Thousands of genes can be measured on a single array, most of which are irrelevant or uninformative for discriminative methods and dimensionality thus must be reduced without losing information.
In this context, our objective was to look for predictors (the genes) that would classify the observed cases (the microarrays) into their known classes. The selection of these

[1]Institut de Mathématiques, Université de Toulouse et CNRS (UMR 5219), F-31062 Toulouse, France
[2]Station d'Amélioration Génétique des Animaux UR 631, Institut National de la Recherche Agronomique, F-31326 Castanet, France
[3]Laboratoire de Génétique Cellulaire UR 444, Institut National de la Recherche Agronomique, F-31326 Castanet, France

discriminative variables can be performed in two ways: either explicitly (filter methods) or implicitly (wrapper methods). The filter methods measure the usefulness of a feature by ordering it with statistical tests such as t- or F-tests. These gene-by-gene approaches are robust against overfitting and computationally fast. However, they disregard the interactions between the features and may fail to find the "useful" set of variables: they usually select variables with redundant information. On the other hand, the aim of the wrapper methods is to measure the usefulness of a subset of features in the set of variables. However, when dealing with a large number of variables as it is the case here, it is computationally impossible to do an exhaustive search among all subsets of features and these methods are prone to overfit. One solution to benefit from the wrapper approach is to perform a search using stochastic approximations that still cover a large portion of the feature space to avoid local minima. The "Optimal Feature Weighting" algorithm (OFW) proposed by Gadat and Younes (2007) allows for the selection of an optimal discriminative subset of variables. This meta algorithm can be applied independently with any classifier. Classifiers such as Support Vector Machines (SVM, Vapnik 1999) and Classification And Regression Trees (CART, Breiman et al. 1984) were passed up to this stochastic meta algorithm in Lê Cao et al. (2007) for 2-class microarray problems. The aim was to make a comparative study of OFW+SVM/CART with other wrapper methods (Recursive Feature Elimination, Guyon et al. 2002, l_0 norm SVM, Weston et al. 2003, Random Forests, Breiman 2001) and the filter method t-test on public microarray data sets. The relevancy of the results was assessed in a statistical manner by measuring the performance of each gene selection, and with a biological expertise related to the biological experiment. The results showed that the selections made with OFW were statistically competitive and biologically relevant, even with complex data sets.

From this point, we investigate this stochastic algorithm with multiclass microarray data sets. Multiclass problems are often considered as an extension of 2-class problems. However this extension is not always straightforward as the data sets are often characterized by unbalanced classes with a small number of cases in at least one of the classes. Furthermore, this "rare" minority class is often the one of interest for the biologists who would like to diagnose a disease for example. Nevertheless, most algorithms do not perform well for such problems as they aim to minimize the overall error rate instead of focusing on the minority class. Moreover, the classification accuracy appears to degrade very quickly as the number of classes increases (Li et al., 2004). Several methods have been proposed in the recent years. Chen et al. (2004) proposed balanced or weighted random forests, McCarthy et al. (2005) compared sampling methods and cost sensitive learning with however no clear winner in the results, and more recently Eitrich et al. (2007); Qiao and Liu (2008) also addressed the unbalanced multiclass issue with cost sensitive machine learning technique or SVM.

In the specific context of multiclass microarray data, Li et al. (2004) applied various

classifiers with various feature selection methods and conclude that the accuracy is highly dependent on the choice of the classifier, rather than the choice of the selection method- although this would be more natural. Chen et al. (2003) applied four filter methods with low correlation between selected genes, Yeung and Burmgarner (2003) applied uncorrelated or error-weighted Shrunken Centroid.
In this study we compare two ways of handling multiclass data: with or without an internal weighting procedure in OFW. We do not intend to optimize the size of the gene subset. We rather focus on the assessment criteria to measure the performance of the different methods on the first selected genes.
Biological interpretation that is one of the main key to evaluate the relevancy of the biological results will not be given in this paper when analyzing the five public data sets, but the reader can refer to Lê Cao et al. (2007) that highlight the importance of biological interpretation in the analysis.
We apply the multicategory classifier CART and the *one-vs-one* SVM approach with OFW on five public microarray data sets. Numerical comparisons are done with Random Forests, known to perform efficiently on such data sets, and one filter method (F-tests), by computing the e.632+ bootstrap error from Efron and Tibshirani (1997) for each feature selection method, the stability of the results with Jaccard Index and by comparing the different gene lists. The weighted and no weighted approaches are then compared in OFW+CART and OFW+SVM with the same tools. Finally, application and biological analysis are performed on a pig folliculogenesis data set.
The first section introduces the theoretical adaptation of the OFW model to the multiclass framework. In next section we consider the computational aspects of the application of CART and SVM in OFW and describe the different tools to assess the performance of the results. Application on public data sets and on a practical data set follow. The paper ends with further elements of discussion.

1 The model

We introduce our model of feature selection in the framework of multiclass analysis. As we focus here on microarray data, we will mostly refer to "genes" instead of "variables".

1.1 Measure of the classification efficiency

Let $\mathcal{G}$ be a large set of genes numbered from 1 to N that describes a signal $\mathcal{I}$ to belonging to one of the classes $\{\mathcal{C}_1, \ldots, \mathcal{C}_k, \ldots, \mathcal{C}_K\}$, $k = 1, .., K$. A classification algorithm $\mathbb{A}$ will be chosen according to the problem type (2-class, multiclass), as OFW does not depend on the classification procedure $\mathbb{A}$.
Let us define a positive weight parameter $\mathbb{P}$ on each of the genes in $\mathcal{G}$. After a normalization step, $\mathbb{P}$ can be considered as a discrete probability on the N genes. The goal

is to learn a probability that fits the efficiency of each gene for the classification of $\mathcal{I}$ in $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$, so that important weights are given to genes with high discriminative power and lower weights to those that have a poorest influence on the classification task. Denote p any small integer compared to N, a gene subset of size p has to be extracted from $\mathcal{G}$ using $\mathbb{P}$. We then define how to measure the goodness of $\mathbb{P}$ for the set of genes $\mathcal{G}$ and the classes $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ (*i.e.* the objective function).

Definition 1 *Given a probability $\mathbb{P}$ on $\mathcal{G}$ and $\epsilon(\omega)$ the measure of classification efficiency with any p-uple $\omega \in \mathcal{G}^p$, the energy of the system at point $\mathbb{P}$ is the mean classification performance where ω is drawn with respect to $\mathbb{P}^{\otimes p}$ in $\mathcal{G}^p$*

$$\mathcal{E}(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[\epsilon] = \sum_{\omega \subset \mathcal{G}^p} \mathbb{P}(w)\epsilon(w). \tag{1}$$

Remark 1 *Remark here that genes selected with respect to $\mathbb{P}$ in (1) are drawn with replacement although it looks more reasonable to use subsets of genes without replacement. This mainly comes from the mathematical derivations to optimize $\mathcal{E}$ that will be described below.*

Note that the energy $\mathcal{E}$ depends on the way we measure the classification efficiency on ω, that we denote $\epsilon(\omega)$. Given any standard classification algorithm $\mathbb{A}$, $\epsilon(\omega)$ will actually be the error rate of $\mathbb{A}$ computed on the training set using the set of extracted features ω. The more $\mathbb{P}$ enables us to hold good genes g for classification (important weight on g and $\epsilon(\omega)$ small each time ω contains this gene g), the less $\mathcal{E}$. Minimizing $\mathcal{E}$ with respect to $\mathbb{P}$ will thus permit to exhibit the most weighted and consequently the most highly discriminative genes. Hence, a natural importance ranking will be read on the weight $\mathbb{P}^\star$ minimizing $\mathcal{E}$.

1.2 Stochastic optimization method

The energy $\mathcal{E}$ can be minimized with a stochastic version of the standard gradient descent technique. More details about the theoretical derivations can be found in Gadat and Younes (2007)
The function $\mathcal{E}$ has to be minimized up to the constraints defined by a discrete probability measure on $\mathcal{G}$. Thus, the more natural way to optimize (1) is to use a gradient descent of $\mathcal{E}$ projected to the set of constraints. The set of constraints $\mathcal{S}$ is the simplex of probability map on $\mathcal{G}$. We also denote by $\Pi_{\mathcal{S}}$ the Affine projection of any point of $\mathbb{R}^N$ on the simplex $\mathcal{S}$. This natural projection $\Pi_{\mathcal{S}}$ of any point x can be computed in a finite number of steps as mentioned in Gadat and Younes (2007). Using this former projection $\Pi_{\mathcal{S}}$, the Euclidean gradient of $\mathcal{E}$ is

$$\forall g \in \mathcal{G} \qquad \nabla\mathcal{E}(\mathbb{P})(g) = \sum_{\omega \in \mathcal{G}^p} \frac{C(\omega, g)\mathbb{P}(\omega)}{\mathbb{P}(g)}\epsilon(\omega), \tag{2}$$

where $C(\omega, g)$ is the number of occurrences of g in ω. The iterative procedure to update $\mathbb{P}$ is then given by

$$\mathbb{P}_{t+dt} = \mathbb{P}_t - \nabla \mathbb{P}_t dt. \tag{3}$$

The main clue is that the Euclidean gradient expression (2) can be seen as an expectation as stated in the next proposition.

Proposition 1 *For any $\mathbb{P}$ probability map on $\mathcal{G}$ and if $\nabla_{\mathcal{S}}$ denotes the gradient of $\mathcal{E}$ with respect to constraints $\mathcal{S}$, $\nabla_{\mathcal{S}}\mathcal{E}$ is given by*

$$\forall g \in \mathcal{G} \qquad \nabla_{\mathcal{S}}\mathcal{E}(\mathbb{P})(g) = \Pi_{\mathcal{S}}\left(\mathbb{E}_{\omega}\left[\frac{C(\omega, g)}{\mathbb{P}(g)}\epsilon(\omega)\right]\right).$$

This last expression is numerically intractable since it requires the computation of ϵ over all possible p-uple of $\mathcal{G}$. To deal with such gradient, a computable Robbins-Monro algorithm can be used, which gets similar asymptotic behavior as (3) (see for instance Gadat and Younes (2007), Kushner and Clark 1978). With this stochastic method, the updated formula of $\mathbb{P}_n$ becomes:

$$\mathbb{P}_{n+1} = \Pi_{\mathcal{S}}\left[\mathbb{P}_n - \alpha_n \frac{C(\omega_n, .)\epsilon(\omega_n)}{\mathbb{P}_n(.)}\right], \tag{4}$$

where ω_n is any set of p genes sampled with respect to $\mathbb{P}_n$. Note that the last expression is always defined since when $\mathbb{P}_n(g) = 0$ as we cannot draw this gene in ω_n and the integer $C(\omega_n, g)$ vanishes. The next theorem precisely describes the asymptotic behavior of (4).

Theorem 1 *Defining the discretisation time $\tau_k = \sum_{i=0}^{k} \alpha_i$ and its associated dual reversion $I(t) = \sup\{k \quad | \quad \tau_k \leq t\}$, then the interpolated process $P^k(t) = \mathbb{P}_{I(\tau_k + t)}$ is an asymptotic pseudo-trajectory of the ordinary differential equation (3) provided that the sequence of steps (α_i) satisfies the two conditions:*

$$\sum_i \alpha_i = \infty \qquad \textit{and} \qquad \exists\, \nu > 0 \quad \sum_i \alpha_i^{1+\nu} < \infty.$$

This last result insures that the stochastic algorithm computing $\mathbb{P}_n$ is asymptotically equivalent to the real gradient descent (3). Several derivations of this theoretical point can be found in Gadat and Younes (2007). In our experiments, we have decided to use a step sequence $\alpha_i = A/(B + i)$ for calibrated constants A and B.

1.3 Detailed algorithm.

We detail the application of the algorithm in the case of a given classifier $\mathbb{A}$:

Let $\mathcal{G} = (\delta_1 \dots \delta_{|\mathcal{G}|})$, $\mu \in \mathbb{N}^*$ and η the stopping criterion.

- For iteration $n = 0$ define $\mathbb{P}_0$ as the uniform distribution on $\mathcal{G}$.
- While $|\mathbb{P}_{(n+\mu)} - \mathbb{P}_n|_\infty > \eta$:
  - extract ω_n from $\mathcal{G}^p$ with respect to $\mathbb{P}_{n,p} = \mathbb{P}_n^{\otimes p}$,
  - construct $\mathbb{A}_{\omega_n}$ and compute $\epsilon(\omega_n)$,
  - compute the drift vector $d_n = C(\omega_n, \cdot)\epsilon(\omega_n)/\mathbb{P}_n(\cdot)$,
  - update $\mathbb{P}_{n+1} = \Pi_{\mathcal{S}}[\mathbb{P}_n - \alpha_n d_n]$,
  - $n = n + 1$.

2 Application of OFW and performance evaluation

We discuss here the applications in the field of multiclass problems. The application of OFW+CART and the comparisons of OFW+CART/SVM in the binary case can be found in Lê Cao et al. (2007).

2.1 CART and SVM multiclass applied to OFW

CART

OFW is applied with the classifier CART (Classification And Regression Trees Breiman et al. 1984) that is well adequate for multiclass problems. CART is constructed via a recursive partitioning routine. It builds a classification rule to predict the class label of the microarrays based on the feature information following the Gini criterion. To avoid overfitting, trees are then generally pruned using a cross validation procedure. In our special case, the trees were not pruned and a node was declared terminal when all the cases landing in this node belonged to the same class.
Note that CART is unstable by nature: a slight change in the features can lead to a very different construction of the tree. Following the example of Breiman (1996), the trees were aggregated (*bagging*) to overcome this instability. As in Breiman (1996), the trees were unpruned, but there is no overfitting, thanks to the aggregation technique. To compute the efficiency criterion ϵ at iteration n we launched B trees on B bootstrap samples on different ω_n^b drawn with respect to $\mathbb{P}_n$, where $b = 1, \dots, B$. We then defined ϵ as the mean classification error rate on the out-of-bag samples. The detailed bagging version of OFW+CART is described in 2.3.

SVM Multiclass

We applied OFW with the *one-vs-one* SVM approach that is implemented in the `e1071` `R` package. Other SVM multiclass approaches could have been applied, such as the *one-vs-rest* approach, the approach proposed by Lee and Lee (2003), by Joachims (1999) or the multiclass version from Weston and Watkins (1999). Unlike CART, SVM is very stable and ϵ was hence computed on only one bootstrap sample ($B = 1$).

2.2 Different computations of the approximate gradient

In contrary to Gadat and Younes (2007), we made some slight modifications of the gradient descent to improve the speed of the algorithm with OFW+CART. We propose an averaged time version of the initial OFW as follows:

$$D_n = \frac{\sum_{i=1}^{n} \alpha_i \bar{d}_i}{\sum_{i=1}^{n} \alpha_i} \quad \text{with} \quad \bar{d}_i = \sum_{b=1}^{B} \frac{C(\omega_i^b, .)\epsilon(\omega_i^b)}{\mathbb{P}_i(.)},$$

where b is the bootstrap sample on which each CART tree is constructed and $\alpha_i = A/(B+i)$ is the step sequenced referred in section 1.2.
This enables the stochastic algorithm to better approximate the mean drift (2) than in the standard case. With CART, the approximation of $\nabla\mathcal{E}$ is actually much more difficult than in the SVM case since the variance of the stochastic algorithm seems higher using CART classifier. This averaging step is hence crucial for the algorithm.

2.3 Detailed OFW+CART algorithm

Here is the detailed version of OFW+CART with bagging.

Let $\mathcal{G} = (\delta_1 \dots \delta_{|\mathcal{G}|})$, $\mu \in \mathbb{N}^*$ and η the stopping criterion. $\mathbb{A}$ is the unpruned classifier CART.

- For iteration $n = 0$ define $\mathbb{P}_0$ as the uniform distribution on $\mathcal{G}$
- While $|\mathbb{P}_{(n+\mu)} - \mathbb{P}_n|_\infty > \eta$:
  - For $b = 1..B$:
    - extract ω_n^b from $\mathcal{G}^p$ with respect to $\mathbb{P}_{n,p} = \mathbb{P}_n^{\otimes p}$,
    - draw a bootstrap sample b_{samp} and construct $\mathbb{A}_{\omega_n^b}^{b_{samp}}$,
    - compute $\epsilon(\omega_n^b)$ on the out-of-bag sample $\bar{b}_{samp}$.
  - compute the averaged drift vector D_n as in 2.2,
  - update $\mathbb{P}_{n+1} = \Pi_{\mathcal{S}}[\mathbb{P}_n - \alpha_n D_n]$,
  - $n = n + 1$.

The last lines introduce a projection Π_S which corresponds to the natural affine projection into the simplex S of discrete probability measures. More precisely, we have

$$\Pi_S(q) = \arg\min_{p \in S} \|q - p\|^2.$$

Note that since $\mathbb{P}_n - \alpha_n D_n$ may have some negative coordinates, this projection is slightly different from a simple normalization step. Several details are provided in Gadat and Younes (2007).

2.4 Weighting procedure

An efficient way to take into account the unbalanced characteristics of the data set is to weight the internal error rate $\epsilon(\omega)$ according to the number samples of each class in the learning set. This would penalize a classification error made on the minority class and hence put more weight on the variables that help classifying this class instead of the majority class.
Let n be the total number of cases and m_k, $k = 1..K$ the number of cases in class k. We define the (normalized) weight of *each* case in class k by $w_k = \frac{1}{m_k \times K}$.
Then for each out-of-bag test case (*i.e.* the sample not drawn in the bootstrap sample), we note mis_k the number of misclassified cases from class k and the weighted internal error rate is defined as:

$$\epsilon(\omega) = \sum_{k=1}^{K} mis_k \times w_k,$$

instead of $\frac{\sum_k^K mis_k}{n}$ in the no weighting case. This weighting procedure also stands for the evaluation step, see following section 2.5.

2.5 Performance measurement

Comparison of the prediction performance

Error rates of all methods on each data set were computed with the e.632+ bootstrap error estimate from Efron and Tibshirani (1997) that is adequate for small sample sizes data sets. Each algorithm will be learned on a bootstrap sample to avoid any overfitting during the gene selection evaluation (see Ambroise and McLachlan 2002). However, note that this performance evaluation does not dictate the optimal number of genes to select. The e.632+ only allows for the comparison of the performances of the different selection methods.

Stability

One can define the feature stability as the level of agreement between the set of selected genes chosen in each bootstrap sample with the set of selected genes using the full training set. The Jaccard index (Yeung and Burmgarner, 2003) then computed lies between 0 (low level of agreement) and 1 (high level of agreement) and will be used to compare the stability of all four ranking methods.

Definition 2 *Let $S(\Delta)$ be the set of the Δ selected genes from the entire training set and $S(nb, \Delta)$ the set of selected genes from the nb bootstrap sample. The number of true positives (TP) is the number of selected genes that were chosen in both $S(\Delta)$ and $S(nb, \Delta)$:*

$$TP = |S(\Delta) \cap S(nb, \Delta)|.$$

Similarly, we define as the false positives (FP) the number of selected genes that were chosen in $S(nb, \Delta)$ but not in $S(\Delta)$:

$$FP = |S(nb, \Delta) \setminus_{S(\Delta)}|,$$

and the number of false negatives (FN) the number of genes that were selected in $S(\Delta)$ but not in $S(nb, \Delta)$:

$$FN = |S(\Delta) \setminus_{S(nb,\Delta)}|.$$

The Jaccard index $J(nb, \Delta)$ is defined as $TP/(TP + FP + FN)$ and is high and close to 1 when there are many true positives and few false positives and false negatives. We then compute the averaged Jaccard index J_Δ over all nb samples for Δ varying between 1 selected gene and Δ_{max} selected genes.

We expect therefore to rank the stability of each feature selection procedure with this Jaccard index.

Table 1: Summary of the five data sets.

	Lymphoma	Leukemia	SRBCT	Brain	Multiple Tumor
# genes	4026	3000[1]	2308	1963[1]	2000[1]
# classes	3	3	4	5	11
# obs.	62	72	63	42	90
# obs. per class	42/9/11	38/9/25	23/20/12/8	10/10/10/4/8	8/4/7/26/ 4/15/3/7/ 6/5/5

[1]pre-filtered with a very large F-test p-value.

2.6 Ranking methods

Multicategory ranking methods are still rare in the context of classification, especially in microarray data context. A comparative study is performed with the well-known Random Forests (RF, Breiman 2001). The three wrapper methods (OFW+CART, OFW+SVM and RF) were also compared to the F-test filter method, that is still widely used for selecting genes in the context of microarrays.
Although Random Forests can also be performed with a weighting approach such as Balanced Random Forests (BRF) or Weighted Random Forests (WRF) from Chen et al. (2004), we chose to compare all these methods with no weighting procedure.

3 Statistical assessment on public data sets

A short description of the five public data sets is first given. We then compare the results obtained with OFW+CART, OFW+SVM, RF and F-test with no weighting procedure. During the evaluation performance, the F-test selection was assessed with a *one-vs-one* linear SVM.
We finally focus on OFW and compare the weighted *vs.* non-weighted procedure and give some elements of discussion.

3.1 Multiclass data sets

We present the results obtained on five public multiclass data sets.

1. Lymphoma (Alizadeh et al., 2000) compares 3 classes of cells (42, 9 and 11 cases per class) with 4026 gene expressions.

2. The 3-class Leukemia version (Golub et al., 1999) with 7129 genes compares the lymphocytes B and T in ALL (Acute Lymphoblastic Leukemia, 38 and 9 cases) and the AML class (Acute Myeloid Leukemia, 25 cases). The classes AML-B and AML-T are known to be biologically very similar.

3. The Small Round Blue-Cell Tumor Data of childhood (SRBCT, Khan et al. 2001) includes 4 different types of tumours with 23, 20, 12 and 8 microarrays per class and 2308 genes.

4. The Brain data set compares 5 embryonal tumours (Pomeroy et al., 2002) with 5597 gene expression. Classes 1, 2 and 3 count 10 microarrays each, the remaining classes 4 and 8.

5. The Multiple Tumor data set initially compared 14 tumors (Ramaswamy et al., 2001) and 7129 gene expressions. We used the normalized data set from Yeung and Burmgarner (2003) with 11 types of tumor. To fit into a usual microarray framework (*i.e.* a small number of samples), we randomly selected 90 samples (out of 192) that have tumor types coming from breast (8), central nervous system (4), colon (7), leukemia (26), lung (4), lymphoma (15), melanoma (3), mesotheolima (7), pancreas (6), renal (5) and uterus (5).

The Brain and the Leukemia data sets were pre-filtered with a very large F-test p-value (0.1 and 0.2, leaving 1963 and 3000 genes). The Multiple Tumor data set was also pre-filtered with an F-test, leaving 2000 genes, to reduce the computation time of the algorithms. These data sets are succinctly described in Table 1.
All these data sets were chosen for their unbalanced characteristics as the minority class represents for each data set a small percentage of the total number of cases.All data sets were assumed to be correctly normalized.

3.2 Comparison of the ranking methods with no weighting procedure

Performance comparison

Figures 1 display the e.632+ error rates obtained on all data sets with respect to the number of selected genes with the different ranking methods.
The classification complexity of the data sets is easy to identify as Lymphoma **(a)** and SRBCT **(c)** display an evaluated error rate less than 7% for a selection of 10 genes, whereas for Leukemia **(b)**, Brain **(d)** and Multiple Tumor **(e)** , the error rates vary between 25 to 50 % for a selection of 10 genes.
OFW is generally among the best performers, and the error rates of OFW+ CART and OFW+SVM are often very close, except for Multiple Tumor, where OFW+SVM gives a poor performance. We suspect that the aggregation of this type of binary SVM (*one-vs-one*) may not be adapted in this extreme multiclass setting.
RF achieves good results on Leukemia, SRBCT and Multiple Tumor, whereas on Lymphoma and Brain, the performance of the RF selection is the worst. RF might therefore not succeed in selecting genes with information relevant enough, especially in Lymphoma, where all classes are easy to classify with too many informative variables. On the contrary, the F-test achieves good results on Lymphoma and Brain. This

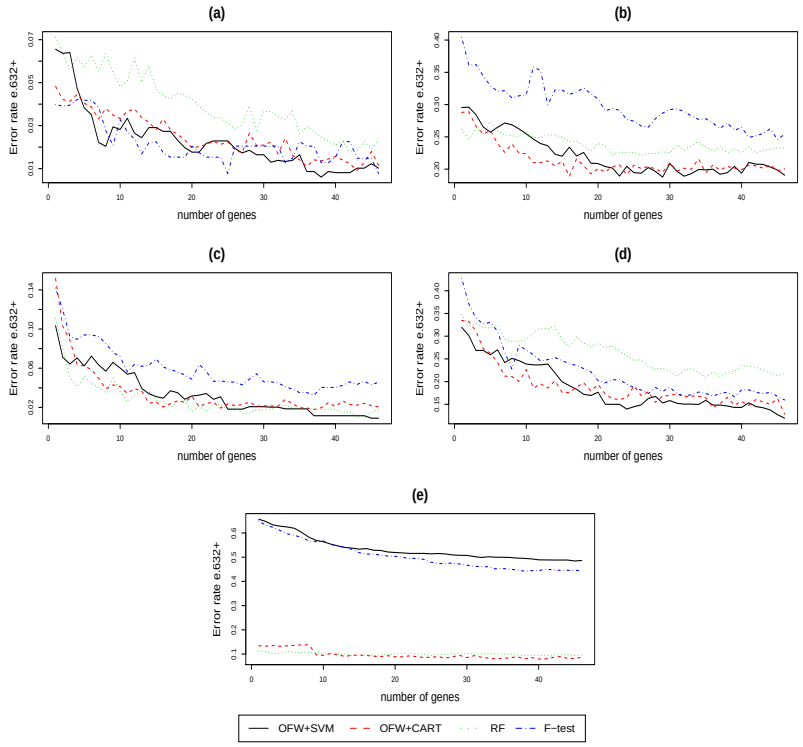


Figure 1: Error e.632+bootstrap of several algorithms with respect to the number of genes on Lymphoma (a), Leukemia (**b**), SRBCT (**c**), Brain (**d**) and Multiple Tumor (**e**) .

filter method orders genes that are differentially expressed (*i.e.* significant) for at least one of the classes. If genes are differentially expressed for more than one class (or for all classes), the selected genes will all be informative enough and the performance will be good. With Leukemia, the F-test performs the worst. This data set is more difficult to classify as the 2 classes ALL-B and AL-LT are very similar (Golub et al., 1999). The difficulty is reinforced as ALL-B is the majority class while ALL-T is the minority class in this 3-class problem. The F-test thus first ordered significant genes that discriminated the easiest class (ALL-B), to the detriment of the other classes.
In any case, these results show that one cannot draw general conclusions on the best method to apply. In general, OFW+SVM and OFW+CART were the best performers, especially OFW+CART in a high multiclass setting.

Remark on the performance assessment with e.632+ bootstrap error rate

The e.632+ error rate was chosen as it is the most adequate to compute the performance of the different methods on small sample data sets (Ambroise and McLachlan, 2002). However we did observe some weaknesses and the interpretation of the results should be done with caution. One would expect the error rate to increase when the number of evaluated variables becomes too big (as more noise enters the selection). This is not the case for any method using the SVM classifier and RF, which are known to base their classification task on the good variables among numerous and possibly noisy variables. The results that we obtain are in agreement with this fact. We did not observe this tendency with OFW+CART, as during the evaluation step, each aggregated tree is constructed on a small variable subset from the selection (see Lê Cao and Chabrier 2008 for the details of the algorithm).
The evaluation error rate should thus be solely used to compare the ranking methods between each others, and not to give an accurate classification error rate of a given variable selection.

Stability

Computation of the Jaccard index with respect to the number of selected genes are displayed in Figures 2. Maximum stability is obtained on easy data sets (Lymphoma **(a)** and SRBCT **(c)**) with a Jaccard index reaching 0.45 and 0.6. The F-test is undoubtedly the most stable method on complex data sets (Leukemia **(b)**, Brain **(d)**, Multiple Tumor **(e)**), although the performance is very poor (see section 3.2). RF is in general very stable compared to OFW+SVM and OFW+CART.
The good stability results of the filter method is easy to explain as the F-test selects redundant information usually only on the majority class, whereas the other methods select genes with relevant information on all classes. As the gene selection might be

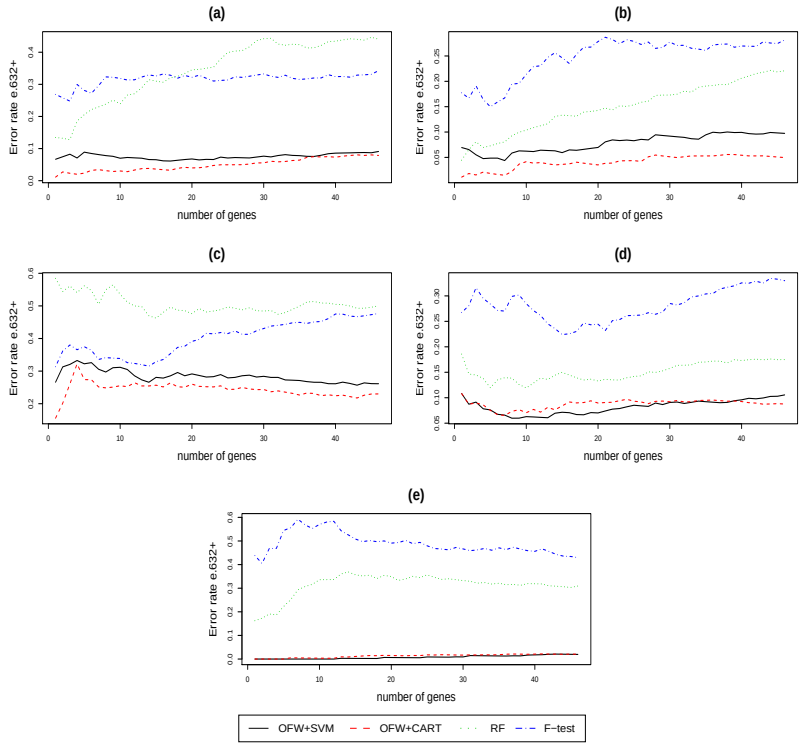


Figure 2: Jaccard index of OFW+SVM, OFW+CART, RF and F-test with respect to the number of genes on Lymphoma (**a**), Leukemia (**b**), SRBCT (**c**), Brain (**d**) and Multiple Tumor (**e**).

strongly dependent on the cases drawn in the bootstrap sample, especially if one of the classes is small, the methods focusing on the minority classes will consequently be less stable.
OFW+SVM and OFW+CART are stochastic methods and are hence less stable for all data sets. When the number of classes becomes large (Brain, SRBCT, Multiple Tumor), the stability results seem largely affected. A compromise needs hence to be taken between information (on all classes) and stability.

Table 2: Number of genes shared by several feature selection algorithms on Leukemia or Lymphoma for a selection of 50 genes.

Leukemia \ Lymphoma	OFW+SVM	OFW+CART	RF	F-test
OFW+SVM	#	12	11	12
OFW+CART	7	#	22	24
RF	17	18	#	30
F-test	3	6	11	#

Table 3: Number of genes shared by several feature selection algorithms on Brain or SRBCT for a selection of 50 genes.

Brain \ SRBCT	OFW+SVM	OFW+CART	RF	F-test
OFW+SVM	#	25	31	11
OFW+CART	8	#	29	15
RF	12	22	#	9
F-test	7	2	2	#

Insight into the different selections

Tables 2 and 3 provide more insight of the different 50 gene lists selected with all methods on each data set (not shown for Multiple Tumor). For example in Table 2 for the Lymphoma data set (upper triangle), OFW+SVM and OFW+CART selected 12 common genes among the 50 selected.
The most striking point is the very few number of shared genes between all methods, that highlights the characteristics of each ranking method. Generally, as they are constructed with the same classifier, RF and OFW+CART share a fair amount of genes (22 and 18 on Lymphoma and Leukemia, Table 2). Table 2 also shows that RF selected more significant genes (*i.e* differentially expressed with F-test) than OFW+CART/SVM (30 and 11 on Lymphoma and Leukemia). In Table 3, where the number of classes is bigger than 3 (SRBCT, Brain), the 3 methods RF, OFW+CART and OFW+SVM generally shared more genes together than with the F-test. This highlights the poor relevancy of a selection made with an F-test in this context.
On all data sets except SRBCT, OFW+CART and OFW+SVM shared very few genes.

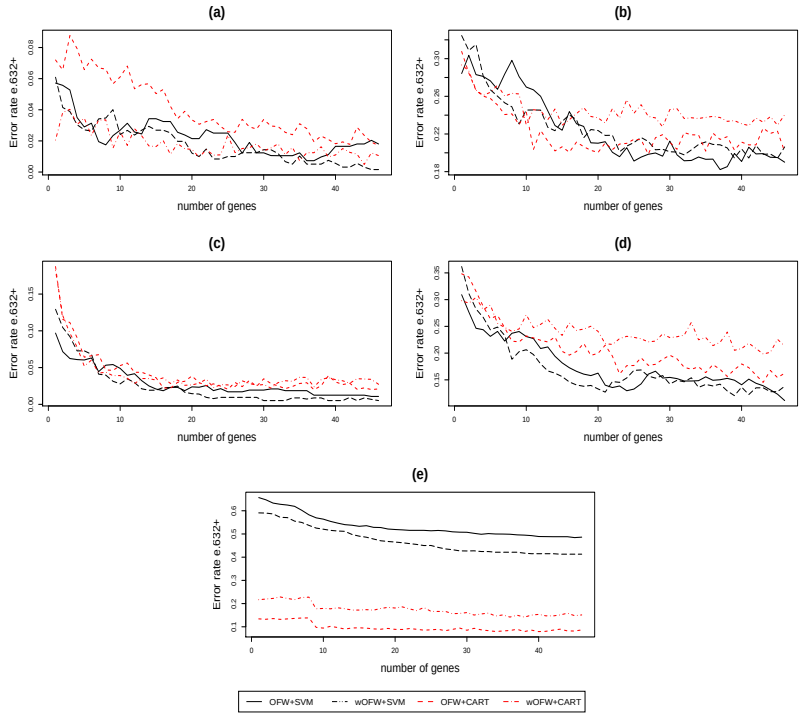


Figure 3: Weighted e.632+ bootstrap error of OFW+CART and OFW+SVM with both procedures weighted and non weighted with respect to the number of genes on Lymphoma (**a**), Leukemia (**b**), SRBCT (**c**), Brain (**d**) and Multiple Tumor (**e**).

This can be explained as the construction of these two classifiers is completely different: CART searches in the feature space the best variable and the best split to divide each node in the tree while SVM looks for the optimal hyperplane between two classes. For SRBCT where all methods except F-test seemed to share numerous genes, this can be explained as all methods seemed to perform equally well with the same relevant genes (see Fig. 1 (**c**)).
Note that the same tendency was observed if we reduced the size of the selection (*e.g.* from 50 to 10): the top selected genes were not necessarily the same from one selection to another.
The difficulty of the Multiple Tumor data set was strongly highlighted as no method shared more than 4 common genes. Given the poor performances of the F-test and OFW+SVM (section 3.2), this small overlapping result is to be expected.

3.3 Comparisons of the weighted and non-weighted procedures of OFW

The aim of this section is to compare the weighted and non-weighted versions of OFW only, as the other ranking methods do not share the same weighting procedure (especially WRF/BRF for RF, Chen et al. 2004), the F-test having no weighting procedure).

Performance comparison

In order to compare the internal weighting procedure in OFW+CART or SVM, we computed the e.632+ error rate for both approaches: weighted (wOFW) or non-weighted (OFW). We remind that the weighted procedure implies an internal weighted error rate in the gradient.
For the e.632+ computations, the learning of the *nb* bootstrap samples of wOFW or OFW for each classifier was performed. Then, during the testing phase, both types of learning were evaluated with a *weighted* e.632+. This was necessary in order to compare the improvement of the performance with the weighting approach. A non-weighting approach in e.632+ would indeed favour the majority class to the detriment of the minority class and would still give a (wrongly) low error rate.
Figures 3 display the weighted e.632+ error rate of OFW and wOFW with the application of either CART or SVM for the five data sets.
There is often a strong difference between the performances of OFW+CART and wOFW+CART, showing that CART seems affected by unbalanced classes, whereas there is no difference between the two variants of OFW+SVM. The *one-vs-one* SVM approach seems hence extremely well adequate for unbalanced classes. wOFW+CART seems to improve the error rate compared to OFW+CART on the easy data set Lymphoma (**a**). For SRBCT **(c)**, all methods perform similarly, whereas for Multiple Tumor **(e)**, wOFW+SVM is still affected by the high number of classes.
These graphs show that the weighting procedure in OFW+SVM seems not necessary in the multiclass case as the *one-vs-one* SVM aims to classify each class, even minority, as long as the number of classes remains reasonable (≤ 5 here). On the contrary, for OFW+CART, the weighting procedure might be needed as by construction, CART tends to favour the majority classes.

Stability

The comparisons of the Jaccard index for both versions of the algorithm is displayed on Figures 4. wOFW+SVM seems to improve the stability of the results of the 3-class data sets Lymphoma (**a**) and Leukemia (**b**). When the number of classes is larger, the non-weighted versions are the most stable.
These Jaccard indexes are very low as the proportion of the minority cases is often

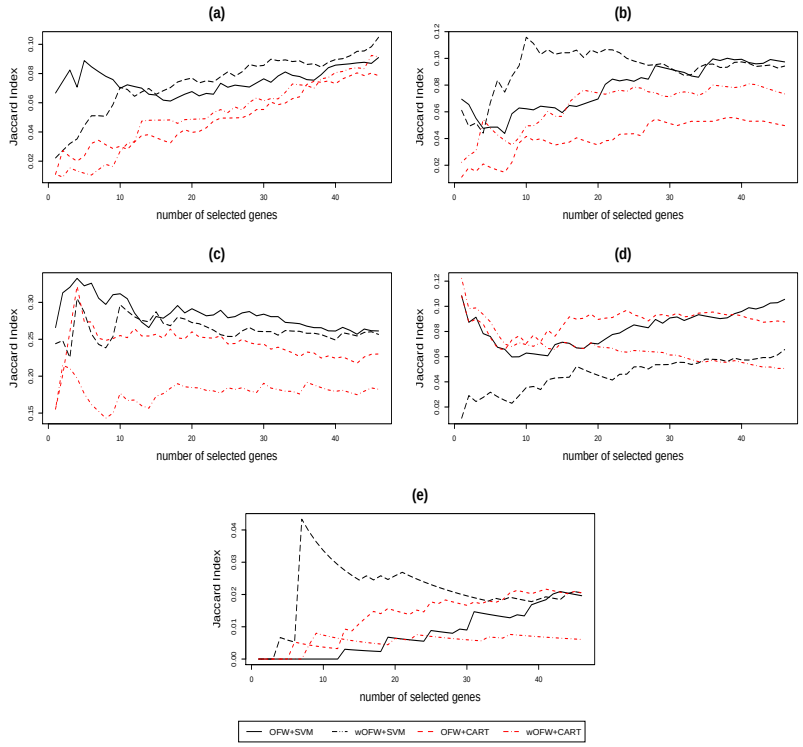


Figure 4: Comparison of the Jaccard index with the weighted and non-weighted versions of OFW+SVM and OFW+CART on Lymphoma (**a**), Leukemia (**b**), SRBCT (**c**), Brain (**d**) and Multiple Tumor (**e**) .

diminished during the bootstrap sampling and the selected variables discriminating the minority classes must strongly depend on each bootstrap sample. This explains the poor results obtained in Multiple Tumor (**e**).

Table 4: Number of genes shared by the weighted and non-weighted versions of OFW+SVM or OFW+CART for each data set (selection of 50 genes).

	Lymphoma	Leukemia	SRBCT	Brain	Multiple Tumor
OFW+SVM ∩ OFW+CART	12	7	29	8	0
wOFW+SVM ∩ wOFW+CART	16	5	24	4	0
OFW+SVM ∩ wOFW+SVM	13	13	31	18	5
OFW+CART ∩ wOFW+CART	27	11	25	13	2

Comparisons of the lists (weighted *vs.* non-weighted)

We compared the lists given by the weighted *vs.* the non-weighted procedures in OFW+CART or SVM in Table 4. There is a difference in the gene selections between the weighted and non-weighted version of OFW. For example on Lymphoma, OFW+SVM and wOFW+SVM shared 13 genes out of the 50 selected. This is surprising as section 3.3 showed that there was not a strong difference in the performance of both methods (Fig. 3 **(a)**). However, with SRBCT, where all performances of the four tested version were similar (Fig. 3 **(c)**), the number of shared genes was quite close and high compared to the other data sets (from 24 to 31 in Table 4).
The less numerous the genes that are shared between OFW and wOFW, the better the improvement of the selection in terms of relevancy (as wOFW aims to favour minority classes). For example the selections of wOFW+SVM in Lymphoma might be more informative than the OFW+SVM selection, the same stands for wOFW+CART *vs.* OFW+CART in Leukemia and Brain. However, the high complexity of the Multiple Tumor data set show the limitation of the algorithm OFW, as well as a strong difference between all proposed versions of this meta algorithm.

4 Application and biological interpretation.

When developing feature selection algorithms for microarray data, we believe it useful to show if the actual gene selection is biologically relevant for the study. The biological interpretation is hence valuable to show the applicability of such algorithms.

4.1 The pig folliculogenesis data set

This experiment was designed to compare different sizes of healthy follicles granulosa cells during the last stages of antral phase. Large (L), Medium-sized (M) and Small (S) follicles from three different sows per size category were used. After extraction,

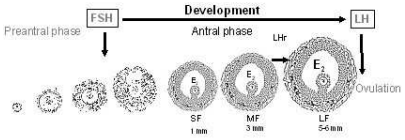


Figure 5: The three follicle classes: Small, Medium-sized and Large.

the RNA isolated from these cells was used to hybridise 42 microarrays that includes duplicates, resulting in 20 Large, 14 Medium-sized and 8 Small follicle cases (GEO accession number: GSE5798). After a normalizing and a filtering steps, the expression of 1564 clones remain on each microarray.

The main characteristic of this data set is the obvious difference between the Large follicles and the others. This is due to the biological properties of the data mainly including the appearance of LH receptors between the Medium and Large follicles (Figure 5). Medium-sized and Small follicles are still in the growth process whereas the Large follicles are completely differentiated to produce steroid hormones. Moreover, during the measurements that assign each follicle its class, the diameters of the Small and the Medium-sized follicles are very similar (1-2mm and 3 mm) whereas the Large ones cannot be mistaken (5-6mm). Another factor to consider is the vast majority of regulated cDNAs (clones) over-expressed in the Large follicles and hence the minority of regulated cDNAs (referred to as *genes* instead of clones) that are over-expressed in the Small ones.
We are clearly here in the practical case where classes are unbalanced, and where the number of original samples is extremely small, as some of the microarray experiments were duplicated.

4.2 Results and biological interpretation

The analysis of this data set with Random Forests and F-test was performed in Bonnet et al. (2008) and gave biologically relevant results. We focus here on the application of OFW+CART/SVM and their weighted variants.

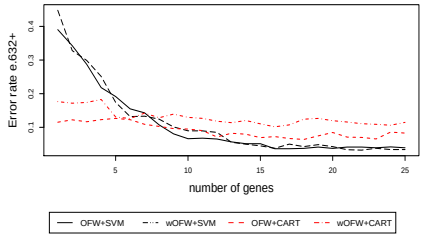


Figure 6: Weighted e.632+ bootstrap error of OFW+CART and OFW+SVM with both procedures weighted and non weighted with respect to the number of genes on the follicle data set.

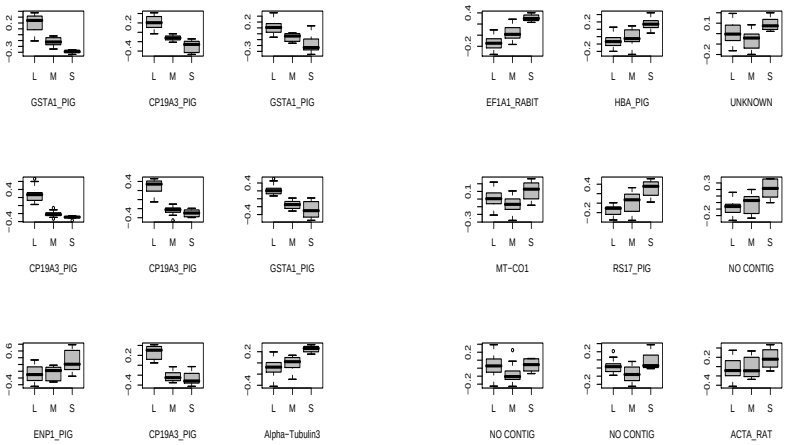


Figure 7: Boxplots of the 9 top genes selection with OFW+CART (left) or with OFW+SVM (right) on the follicle growth data set. Boxplots are displayed for each class (L, M and S).

Application of OFW

When the number of original samples is extremely small, the e.632+ bootstrap error rate must be considered with caution and should not be the only argument to favour a gene selection coming from a feature selection method rather than another. Fig. 6 displays the weighted e.632+ error rate for all approaches. Both OFW+SVM and wOFW+SVM seem to give the best performance.
However, our experience show that the most biologically relevant results do not always give the best statistical performance (Lê Cao et al., 2007). This is why biological interpretation is a crucial step when analyzing microarray data.

Interpretation of the results

In these four gene lists we identified the genes GSTA1 and Cyp19A3 which are known to be over-expressed during follicular development (Keira et al., 1994; Slomczynska et al., 2003) and nexin, ACTA2, ATF7, UBC, that were not selected by F-test and Random Forest in the previous analysis.
Figure 7 displays the boxplots of the 9 top genes selected either with OFW+CART or OFW+SVM for each class (L, M or S). They show that while a minority of selected genes are over-expressed in the S class with OFW+CART (left), a majority of them are over-expressed in the S class in the OFW+SVM selection (right). This tendency can be generalized for a larger list of genes. It seems here that the construction of the *one-vs-one* SVM tends to mostly favour genes discriminating the minority class S rather than the majority class L, as L seems too easy to classify.
When applying wOFW+CART and wOFW+SVM, this tendency is still observed, with more genes that are over-expressed in S for the wOFW+CART selection (not shown). The biological analysis shows that most of the over-expressed genes in the S class code for ribosomic proteins that may be associated with a decrease of proliferation during follicular growth from Small to Medium follicles. The wOFW+SVM selection seems hence to give a better discrimination between S and M classes. However, we also identify in this selection a great number of unknown genes that will need further investigation. The wOFW+CART selection seemed not appropriate here since two negative controls were selected and the OFW+SVM selection missed the known discriminative gene CYP11A3.
This section shows that depending on the experimental design, as well as the precise biological questions, the statistician might not answer the study's aim if the conclusions are only drawn from statistical results.

5 General remarks

5.1 Computation time.

The experiments were performed with `R` with a 1.6 GHz 960 Mo RAM AMD Turion 64 X2 PC for OFW+SVM (implementation in `R`) and OFW+CART (implementation in C in a `R` package). The learning time of OFW mostly depends on the initial number of variables in the feature space and the step of the stochastic scheme, as well as the size of ω and the number of trees aggregated for OFW+CART. For Brain (Lymphoma) that contains 1963 (4026) genes, the learning took about 1 (1.5) hour for OFW+SVM for 200 000 iterations.It took 1 (3.5) hour for OFW+CART for 5000 iterations.

5.2 Complexity of OFW.

The complexity of the meta algorithm OFW depends on two points. The first one is the nature of the algorithm used with SVM. The second point is the convergence speed of the stochastic scheme towards a minimum of the energy $\mathcal{E}$.

The complexity of each algorithm used with OFW (CART, SVM, Multiclass SVM, ...) may be very variable and depends on the choice of the user. For instance, with this meta algorithm, each iteration computes a SVM with N_s samples described by p variables and the complexity of each step is at most $p \times N_s^2$ since $p > N_s$ in this study (see detailed computation of this complexity in Burges 1998).

Regarding the second points, the convergence to an optimal state x^* using a standard (non averaged) Robbins-Monro stochastic approximation scheme $(X_n)_{n\in\mathbb{N}}$ is described by the following assessment:

$$\sqrt{\frac{n}{\log n}}(X_n - x^\star) \to \mathcal{N}(0, \Lambda^*). \tag{5}$$

This last theoretical derivation can be found in Duflo (1997). In this last statement, Λ^* is the trace of Hessian matrix of $\mathcal{E}$ computed on the optimal state x^*. If n iterations are run in the initial version of OFW Gadat and Younes (2007), the convergence speed is bounded by $O\left(\frac{\log n}{n} Tr(\Lambda^*)\right)$. The interest of the OFW meta algorithm is significant since an exhaustive search of p-uple among N features would required C_N^p iterations.

The interest of the averaging step introduced in section 2.2 is to improve the rate of convergence of the stochastic scheme reducing the variance of the estimate D_n. The theoretical derivations concerning the rate of convergence is at the moment an open issue but it is likely to reduce the $Tr(\Lambda^*)$ term introduced in (5).

5.3 General remarks

This study shows that microarray data sets have various levels of difficulty and are quite unpredictable if there is not a solid biological knowledge background of the data set. The analysis of several public data set shows that there is no data set that seems to behave like the other. Without biological expertise, it is extremely difficult to assess the relevancy of the results. Simulating a set of data would not help giving more insight in the applied methodologies, as simulating a data set like microarray is an extremely complex work.

The performance assessment of the methods could be computed, but had sometimes serious limits, due to the evaluation method and the applied algorithms, or the small number of samples. This study shows that the evaluation part has to be taken with caution by the user in search of the "best" method.

Furthermore, although there seemed to be no improvement of the performance of the method when applying wOFW+SVM, the resulting gene selection seemed to contain

more biological information on the minority class. Our evaluation performance method might hence not be adequate in this context, especially for OFW+CART where a "double bootstrap sampling" is performed during the evaluation step. We also believe that the performance of wOFW+CART can be improved by directly including weights during the construction of the trees.
Both multicategory classifiers CART and *one-vs-one* SVM that were applied with OFW seemed to perform better than the other tested methods, except when the number of classes was very high (here ≥ 5). In this case, aggregating binary *one-vs-one* SVMs seems limited. Lee and Lee (2003) mentioned that the *one-vs-rest* SVM can also give bad results if several classes are similar, as it is often the case with biological data. One should investigate instead the implementation of a multiclass SVM, as was proposed by Weston and Watkins (1999), to solve the multiclass optimization quadratic problem into the SVM directly rather than aggregating binary SVMs.
Regarding the performances, choosing between these two methods seems difficult. If the user is interested in biological relevancy of the gene selection, or if the number of classes is high, then OFW+CART might be adequate as the construction of CART really fits this requirement (*i.e* finding genes with differential expression in different classes at each node of the tree). However if the interest mostly lies in the classification task and finding predictive genes, then OFW+SVM might be appropriate. By construction, it searches the best hyperplane between two of the classes. In contrary to CART, SVM optimizes a cost criterion based on the classification performance.

6 Conclusion

Starting from Lê Cao et al. (2007) that provided interesting results for binary problems, we extended the application of OFW+CART and OFW+SVM *one-vs-one* for multiclass microarray problems. These data sets are known to be difficult because of their high dimensionality with a small sample size and at least one of the classes that is under represented. For most classifiers, this often results in a good overall classification accuracy even though the minority classes are misclassified.
We first compared OFW+CART and OFW+SVM with two other methods, Random Forests and the still widely used F-test in gene selection. All methods were performed with no weighting procedure. Our results showed that our two methods generally gave good results in terms of error rate estimation. The filter method F-test seemed not appropriate for multiclass datasets and the stability of the results tended to be better in OFW+SVM than CART.
We then compared the weighted version of wOFW+CART or SVM. There seemed to be no difference in the performance evaluation between the weighted and the non-weighted version of OFW+SVM, which generally performed the best. The performances of the two versions of OFW+CART differed largely, due to the extensive use of bootstrap

samples during the learning step. The relevancy of the selected genes with wOFW should however be improved as they aim at discriminating the minority classes.
In the case where the classes were numerous (≥ 5) and unbalanced, OFW+CART clearly outperformed OFW+SVM. These poor results were due to the type of binary SVMs that were aggregated for the multiclass purpose. The implementation of OFW with a multiclass SVM might improve these results.
Application and biological interpretation on a real world data set (pig folliculogenesis data set) show that the wOFW+SVM selection might give relevant results that are complementary with a previous analysis.

Availability

OFW is implemented in an `R` package called `ofw`.

References

Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403:503–511.

Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction in tumour classification on basis of microarray gene expression data. Proc. Natl. Acad. Sci. USA, 99(1):6562–6566.

Bonnet, A., Lê Cao, K., SanCristobal, M., Benne, F., Tosser-Klopp, G., Robert-Granié, C., Law-So, G., Besse, P., De Billy, E., Quesnel, H., et al. (2008). Identification of gene networks involved in antral follicular development. Reproduction.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2):123–140.

Breiman, L. (2001). Random forests. Machine Learning, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2:121–167.

Chen, C., Liaw, A., and Breiman, L. (2004). Using random forest to learn imbalanced data. Technical Report 666, Dpt. of Statistics, University of Berkeley.

Chen, D., Hua, D., Reifman, J., and Cheng, X. (2003). Gene selection for multi-class prediction of microarray data. In CSB '03: Proceedings of the IEEE Computer Society Conference on Bioinformatics, page 492, Washington, DC, USA. IEEE Computer Society.

Duflo, M. (1997). Random Iterative Models. Springer.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the e.632+ bootstrap method. Journal of American Statistical Association, 92:548–560.

Eitrich, T., Kless, A., Druska, C., Meyer, W., and Grotendorst, J. (2007). Classification of Highly Unbalanced CYP450 Data of Drugs Using Cost Sensitive Machine Learning Techniques. Journal of Chemical Information and Modeling, 47(1):92–103.

Gadat, S. and Younes, L. (2007). A stochastic algorithm for feature selection in pattern recognition. J. Mach. Learn. Res., 8:509–547.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., et al. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 286(5439):531.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 46(1):389–422.

Joachims, T. (1999). Making large-Scale SVM Learning Practical. MIT-Press.

Keira, M., Nihihira, J., Ishibasshi, T., Tanaka, T., and Fujimoto, S. (1994). Identification of a molecular species in porcine ovarian luteal glutathione S-transferase and its hormonal regulation by pituitary gonadotropins. Archives of biochemistry and biophysics(Print), 308(1):126–132.

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med, 7(6):673–679.

Kushner, H. and Clark, D. (1978). Stochastic Approximation Methods for Constrained and Unconstrained Systems. Springer-Verlag.

Lê Cao, K.-A. and Chabrier, P. (2008). ofw: an R package to select continuous variables for multiclass classification with a stochastic wrapper method. Journal of Statistical Software (in press).

Lê Cao, K.-A., Gonçalves, O., Besse, P., and Gadat, S. (2007). Selection of biologically relevant genes with a wrapper stochastic algorithm. Statistical Applications in Genetics and Molecular Biology, 6(1):Article 1.

Lee, Y. and Lee, C. (2003). Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics, 19(9):1132–1139.

Li, T., Zhang, C., and Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20(15):2429–2437.

McCarthy, K., Zabar, B., and Weiss, G. (2005). Does cost-sensitive learning beat sampling for classifying rare classes? Proceedings of the 1st international workshop on Utility-based data mining, pages 69–77.

Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C., et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature, 415(6870):436–442.

Qiao, X. and Liu, Y. (2008). Adaptive Weighted Learning for Unbalanced Multicategory Classification. Biometrics, (0).

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences, 98(26):15149–15154.

Slomczynska, M., Szoltys, M., Duda, M., Sikora, K., and Tabarowski, Z. (2003). Androgens and FSH affect androgen receptor and aromatase distribution in theporcine ovary. Folia Biol (Krakow), 51:63–68.

Vapnik, V. N. (1999). The Nature of Statistical Learning Theory (Information Science and Statistics). Springer.

Weston, J., Elisseeff, A., Schölkopf, B., and Tipping, M. (2003). Use of the zero norm with linear models and kernel methods. The Journal of Machine Learning Research, 3:1439–1461.

Weston, J. and Watkins, C. (1999). Support vector machines for multi-class pattern recognition. Proceedings of the Seventh European Symposium On Artificial Neural Networks, 4:6.

Yeung, K. and Burmgarner, R. (2003). Multi-class classification of microarray data with repeated measurements: application to cancer. Genome Biology, 4(83).

5. Logiciel développé

Le développement de l'algorithme OFW appliqué à SVM et CART et sa mise à disposition dans une librairie `R` a pris une place importante dans le travail de la thèse afin de pouvoir valoriser l'approche developpée et la rendre accessible. Cet algorithme de type *wrapper* est coûteux en temps de calcul et a demandé lors de sa phase de mise au point une implémentation sur une machine de calcul parallèle, avant d'être ré-écrit en langage `C` et `R` pour économiser considérablement le temps de calcul.
Ce package `R` est maintenant accessible sur la page du CRAN, et le tutoriel que nous présentons a pour but de faciliter l'usage de cette approche. Un code de `R` parallèle, avec la librairie `Rmpi` est aussi proposé pour réduire le temps de calcul lors de l'étape d'évaluation.

Cet article a été accepté dans la revue Journal of Statistical Software (accepté le 25 juillet 2008, sous presse).

ofw: an R package to select continuous variables for multiclass classification with a stochastic wrapper method

Kim-Anh Lê Cao[1,2] and Patrick Chabrier[3]

Abstract

When dealing with high dimensional and low sample size data, feature selection is often needed to help reduce the dimension of the variable space while optimizing the classification task. Few tools exist for selecting variables in such data sets, especially when classes are numerous (> 2).
We have developed `ofw`, an `R` package that implements, in the context of classification, the meta algorithm "Optimal Feature Weighting" (OFW). We focus on microarray data, although the method can be applied to any $p >> n$ problems with continuous variables. The aim is to select relevant variables and to numerically evaluate the resulting variable selection. Two versions of OFW are proposed with the application of supervised multiclass classifiers such as CART and SVM. Furthermore, a weighted approach can be chosen to deal with unbalanced multiclasses, a common characteristic in microarray data sets.
`ofw` is freely available as an `R` package under the GPL license. The package can be downloaded from the Comprehensive R Archive Network (CRAN).

Introduction

Performing a feature selection algorithm has several important applications in high dimensional data sets. For example with microarray data, it is sensible to use a dimensional reduction technique, either to identify genes that contribute the most for the biological outcome (e.g., cancerous vs. normal cells) and to determine in which way they interact to determine the outcome, or to predict the outcome when a new observation is presented. Such a method would provide practical aspects with machine learning methods: it avoids the "curse of dimensionality" that leads to overfitting when the number of variables is too large.

[1]Institut de Mathématiques, Université de Toulouse et CNRS (UMR 5219), F-31062 Toulouse, France
[2]Station d'Amélioration Génétique des Animaux UR 631, Institut National de la Recherche Agronomique, F-31326 Castanet, France
[3]Biométrie et Intelligence Artificielle, UR875, Institut National de la Recherche Agronomique, F-31326 Castanet, France

There are two ways of selecting features. Either explicitly (filter methods) or implicitly (wrapper methods). The filter methods measure the relevance of a feature at a time by performing statistical tests (e.g., t test, F test) and ordering the p values. This type of approach is robust against overfitting and is fast to compute. However, it usually disregards the interactions between the features as it tests one variable at a time. Chen et al. (2003) compared four filter methods and reached this conclusion.

The wrapper methods measure the usefulness of a feature subset by searching the space of all possible feature subsets. The search can be performed either with heuristic or stochastic search. The main disadvantages of these methods are their tendency to overfit and when dealing with numerous variables, an exhaustive search is computationally impossible. However, the resulting selection takes into account the interactions between variables and might highlight useful information on the experiment. Despite this latter property, wrapper methods are still not widely applied in microarray data. Comparisons of Random Forests (Breiman, 2001), Recursive Feature Elimination (Guyon et al., 2002), L_0 norm SVM (Weston et al., 2003) and biological interpretation of the resulting gene selections is given in Lê Cao et al. (2007b).

In this `R` package, we implement the wrapper method "optimal oeature weighting" (OFW) adapted from Gadat and Younes (2007) that numerically quantifies the classification efficiency of each variable with a probability weight, by using stochastic approximations. This meta algorithm can be applied to any classifier. Therefore, the classifiers SVM (support vector machines, Vapnik 1999) and CART (classification and regression trees, Breiman et al. 1984) have been implemented so as to select an optimal subset of dicriminative variables. Few wrapper methods have been proposed yet to deal with multiclass data sets (Li et al., 2004; Chen et al., 2003; Yeung and Burmgarner, 2003), especially when the classes are unbalanced (Chen et al., 2004). Our function `ofw()` proposes a weighting approach to deal with this common characteristic in microarrays.

Furthermore, like any wrapper methods, `ofw` requires heavy computations, especially when the number of variables is large. In this package, some of the computation time has been reduced by implementing some `C` functions and by proposing parallel programming during the learning step.

Finally, we propose to perform the e.632+ bootstrap method (Efron and Tibshirani, 1997) to estimate the classification error rate on bootstrap samples and to evaluate the different variants of OFW and the resulting gene selections.

The general principle of the OFW algorithm is first presented. We then detail how to use `ofw` by applying the main functions on one microarray data set that is available in the package.

1 Optimal Feature Weighting model

1.1 Principle

OFW (Gadat and Younes, 2007)) is a meta algorithm that can treat several classification problems with a feature selection task. Any classifier can be applied, and Lê Cao et al. (2007a) implemented OFW with CART and SVM for multiclass classification (see also Lê Cao et al. 2007b for binary case).

We assume that the n examples (or cases) are described by p attributes (or variables) and labelled with their target class (e.g., $\{0, 1\}$ in binary problems).

Given a probability weight vector $\mathbb{P}$ on all p variables, the key idea of OFW is to learn $\mathbb{P}$ such that it fits the classification efficiency of each variable in the given problem. In short, important weights will be given to variables with a high discriminative power, and low or zero weights to non relevant variables in the classification task.

For that purpose, the algorithm adopts a wrapper technique, by drawing a small variable subset ω at a time, by measuring the relevance of this subset with the computation of the classification error rate, and then by updating the probability weights $\mathbb{P}$ according to the discriminative power of the variable subset ω. As an exhaustive search of the whole variable space is not tractable when p is large (in microarray data $p > 5000$), stochastic approximations are proposed, see Gadat and Younes (2007); Lê Cao et al. (2007b) for the detailed theory of the model. At iteration i in the algorithm, the probability weight vector is updated with a gradient descent:

$$\mathbb{P}_{i+1} = \Pi_{\mathcal{S}}[\mathbb{P}_i - \alpha_i d_i]$$

where $\Pi_{\mathcal{S}}$ is the projection on the simplex of probability map on the set of variables, so that $\mathbb{P}_{i+1}$ remains a probability vector, α_i is the step of the gradient, and d_i is the stochastic approximation of the gradient (see below).

The whole process is repeated *iter.max* times and the final output is $\mathbb{P}_{iter.max}$, that indicates the importance of each variable in the data. To obtain a variable selection, the user only needs to rank the variables according to their decreasing weights, and to choose the length of the selection.

1.2 General algorithm

Input: a data matrix of size $n \times p$ and the class values vector of size n.

Parameters: number of total iterations *iter.max* and the size `mtry` of the variable subset ω.

Output: $\mathbb{P}_{iter.max}$ a weight vector of length p.

Initialize $\mathbb{P}_0 = [1/p, \ldots, 1/p]$ (uniform distribution on all variables)

For i= 1 to *iter.max*

1. Variables: draw a subset ω_i with respect to $\mathbb{P}_i$

2. Cases: draw a bootstrap sample B_i in $1, \ldots, n$ and define $\bar{B}_i$ the out-of-bag cases
3. Train the classifier on variables in ω_i and cases in B_i
4. Test the classifier on variables in ω_i and cases in $\bar{B}_i$, compute the classification error rate ϵ_i
5. Compute the drift vector d_i
6. Update $\mathbb{P}_{i+1} = \Pi_{\mathcal{S}}[\mathbb{P}_i - \alpha_i d_i]$

where for each iteration i:

- $d_i = \frac{C(\omega_i,.)\epsilon_i}{\mathbb{P}_i(.)}$ is the approximated gradient, and $C(\omega_i, k)$ is the number of occurrences of variable k in the subset ω_i, in case this variable is drawn more than one time in ω_i.
- $\Pi_{\mathcal{S}}$ is the projection on the simplex, so that $\sum_{j=1}^{p} \mathbb{P}_{ij} = 1$ and $\forall j \quad \mathbb{P}_{ij} \geq 0$, $j = 1, \ldots, p$.
- α_i is the step of the gradient descent, and is set to $\frac{1}{i+10}$.

1.3 OFW is applied with either CART or SVM

We applied OFW with two supervised algorithms: SVM and CART.

Support vector machines

SVM SVM (Vapnik, 1999) is a binary classifier that attempts to separate the cases by defining an optimal hyperplane between the 2 classes up to a consistency criterion. Linear kernel SVMs are performed here because of their good generalization ability compared to more complex kernels.

SVM for multiclass data We applied OFW with the one-vs.-one SVM approach that is implemented in the `e1071` `R` package. `ofw` hence depends on `e1071`. The user only needs to set the total number of iterations to perform (`nsvm`) and the size `mtry` of the subset ω to draw at each iteration (see Section 3.2 for tuning).

Classification and regression trees

OFW is applied with the multiclass classifier CART (Breiman et al., 1984) that is well adequate for multiclass problems. Following the example of Breiman (1996), the trees were aggregated (*bagging*) to overcome their unstable characteristic. Hence, several classification trees are constructed on different bootstrap samples and with

different subsets ω. The approximated gradient is also slightly modified. The modified algorithm is as follows:

Input: data matrix of size $n \times p$ and the class values vector of size n.

Parameters: number of total iterations *iter.max*, the size `mtry` of the variable subset ω and the number `ntree` of trees to aggregate.

Output: $\mathbb{P}_{iter.max}$ a weight vector of length p.

Initialize $\mathbb{P}_0 = [1/p, \ldots, 1/p]$ (uniform distribution on all variables)

For $i = 1$ to *iter.max*

1. **For** $b = 1$ to *ntree*
   (a) Variables: draw a subset ω_i^b with respect to $\mathbb{P}_i$
   (b) Cases: draw a bootstrap sample B_i^b in $1, \ldots, n$ and define $\bar{B}_i^b$ the out-of-bag cases
   (c) Train the classifier on variables in ω_i^b and cases in B_i^b
   (d) Test the classifier on variables in ω_i^b and cases in $\bar{B}_i^b$, compute the classification error rate ϵ_i^b
2. Compute the drift vector D_i
3. Update $\mathbb{P}_{i+1} = \Pi_{\mathcal{S}}[\mathbb{P}_i - \alpha_i D_i]$

where D_i is an averaged time version of the gradient d_i (see Lê Cao et al., 2007b).

Hence, as in Random Forests (Breiman, 2001), `ntree` trees are constructed on `ntree` bootstrap samples. The only difference lies in the construction of the classification trees: instead of randomly selecting a variable subset to split each node of each tree (Random Forests), the variable subset is drawn with respect to the probability $\mathbb{P}_i$ to construct each tree.

In addition to choose the total number of iterations to perform (`nforest`) and the size `mtry` of the subset ω to draw at each iteration, the user needs to choose the number of aggregated trees `ntree` (see Section 3.2 for tuning).

1.4 Unbalanced Multiclass

Challenge when data are unbalanced

Multiclass problems are often considered as an extension of 2-class problems. However this extension is not always straightforward, especially in microarray data context. Indeed, the data sets are often characterized by unbalanced classes with a small number of cases in at least one of the classes. This imbalance is often due to rare classes (e.g., a rare disease where patients are few) that are biologically interesting. Nevertheless, most algorithms do not perform well for such problems as they aim to minimize the overall error rate instead of focusing on the minority class.

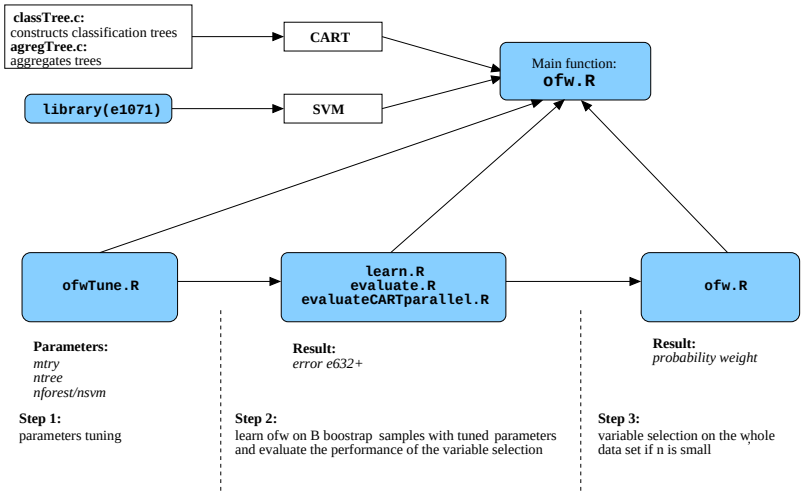


Figure 1: Schematic view of the data set analysis with ofw. The user only needs to use the R functions (in blue).

Weighted procedure in OFW: wOFW

An efficient way to take into account the unbalanced characteristics of the data set is to weight the error rate ϵ_i according to the number samples of each class in the bootstrap sample. This allows for penalizing a classification error made on the minority class and, therefore, put more weight on the variables that help classify this latter class instead of the majority one (Lê Cao et al., 2007a).

This weighted approach has been implemented in both versions of the algorithm, called ofwCART and ofwSVM, and also stands for the evaluation step (**step 2** in Figure 1 and see Section 3.4).

2 Implementation issues

`ofw` is available from the Comprehensive `R` Archive Network (CRAN, `http://CRAN.R-projec` or one of its mirrors). Instructions for package installation are given by typing `help("instal` in `R`.

`ofw` is a set of `R` and `C` functions to perform either ofwCART or ofwSVM and to evaluate the performances of both algorithms. Two classes of functions in `R` and `C` are implemented. Figure 1 provides a schematic view of the analysis of a data set with `ofw`. Each step in Figure 1 will be detailed in Section 3 on a small microarray data set.

The R environment is the only user interface. The R procedure calls a C subroutine, whose results are returned to R. There is no formula interface and the predictors can be specified as a matrix or a data frame via the x argument, with factor responses as a vector via the y argument. Note that ofw performs only classification and does not handle categorical variables. Details of the components of each object from ofwTune, ofw, learn, evaluate and evaluate CARTparallel are provided in the online documentation. Methods provided for the classes ofwTune and ofw include print.

The C function classTree.c that constructs classification trees has been borrowed from the Breiman and Cutler's Fortran programs and converted to C language. The function agregTree.c that aggregates trees was then largely inspired from the randomForest package (Liaw and Wiener, 2002).

3 Using ofw

We detail the call to functions and R commands (preceded by the prompt symbol R>) of ofw, that can be loaded into R by R> library("ofw").

3.1 Illustrative data set

ofw was previously tested on several published miroarray data sets (Lê Cao et al., 2007b,a) by comparing it with several other wrapper algorithms. We comment on the present paper the results obtained on one data set that is provided as an example in the package. SRBCT (Khan et al., 2001) is the data set of small round blue cell tumors of childhood. The training set consists of 63 training samples spanning 4 classes. The data set available in the package includes 2308 genes out of the 6567 after filtering for a minimal level of expression (performed by Khan et al. 2001). Further details about this data set can be found in http://research.nhgri.nih.gov/microarray/Supplement. In order to minimize the computation time in this illustrative example, we have reduced SRBCT to 200 genes by simply randomly selecting these out of the 2308 in the initial data set. We also added a factor class that indicates the class of each microarray sample. Note that normalization of the data, that is a crucial step in the analysis of microarray data is not dealt with ofw and has to be performed first by the user.

3.2 Tuning parameters

In the algorithm OFW, there are mainly 2 to 3 parameters to tune according to the applied classifier to ensure that OFW converges (**step 1** in Figure 1):

1. the size of the gene subset ω (called mtry).
2. the total number of iterations (called nsvm for ofwSVM and nforest for ofwCART).

3. the number of trees `ntree` to agregate for ofwCART.

The package `ofw` provides the function `ofwTune` to tune these parameters. Here is the command to launch `ofwTune` with ofwCART for different `mtry` values:

```
R> data("srbct")
R> attach(srbct)
R> tune.cart <- ofwTune(srbct, as.factor(class), type="CART", ntree=150,
+ nforest=3000, mtry.test=seq(5,25,length=5), do.trace=100, nstable=25)
R> detach("srbct")
```

Note that the only arbitrary parameter that is not tuned and has to be provided by the user is the number of variables `nstable` one wants to select (see below).

Tuning mtry. The function `ofwTune` consists in testing OFW (with CART or SVM) with several sizes of the subset ω (`mtry.test}). Then, for each \verb`mtry.test?, OFW is performed twice, called `ofw1` and `ofw2`. The first `nstable` variables with the highest weights in $\mathbb{P}^{\texttt{ofw1}}_{\texttt{nforest}}$ and $\mathbb{P}^{\texttt{ofw2}}_{\texttt{nforest}}$ are extracted. The `ofwTune` function then outputs the intersection length of these two variable selections. For example, to tune the parameters with ofwCART:

```
R> tune.cart$param
        1  2  3  4  5
mtry    5 10 15 20 25
length 13  9  9  7  5
```

This outputs the intersection length of the first `nstable` variables for each tested `mtry.test`. The value `mtry`$= 5$ gets the best stable results and should be chosen for **steps 2** and **3** in Figure 1 (evaluation and variable selection steps).

Early stopping. Instead of running OFW for all iterations, the user can choose instead to set the number of variables (`nstable`) to select in the final variable selection step (**step 3**). This halts the algorithm once it becomes "stable", that is, when the `nstable` features of highest weights in $\mathbb{P}_i$ and $\mathbb{P}_{i+\texttt{do.trace}}$ are the same for iterations i and $i +$ `do.trace`.

Finally, to choose the total number of iterations in **step 3**, we simply suggest to take 2 to 3 times the number of iterations that were performed using the early stopping criterion, to ensure the convergence of the algorithm. This command outputs the number of iterations which were performed:

```
R> tune.cart$itermax
           1   2   3   4   5
ofwCART1 700 500 900 100 100
ofwCART2 800 700 500 800 800
```

Here the two algorithms `ofwCART1` and `ofwCART2` stopped at 700 and 800 iterations for `mtry=5`. During the final learning step, the user should hence set `nforest`$= 3 * 800$.

Tuning ntree (ofwCART). The best way to tune `ntree` would be then to run `ofwTune` with different values of `ntree` and choose the one that gets the largest intersection length of the first `nstable` variables. In our experience, the more numerous the trees, the more stable the results, usually for `ntree=100` to `150`. The same stands for the weighted (`weight=T`) or non-weighted (`weight=F`) versions of OFW.

An example with ofwSVM. With the SVM classifier, the user has to specify `type="SVM"` and use `nsvm` instead of `nforest` to indicate the number of chosen iterations. As SVM are not aggregated, the user should set `nsvm` >> `nforest`.

```
R> tune.svm <- ofwTune(data, as.factor(class), type="SVM", nsvm=200000, mtry
+ mtry.test=seq(5,25,length=5), do.trace=2000, nstable=25)
R> tune.svm$param

       1  2  3  4  5
mtry   5 10 15 20 25
length 7  6  6  1  2

R> tune.svm$itermax

            1     2     3    4     5
ofwSVM1  8000  4000  8000 4000  4000
ofwSVM2 10000 10000 12000 4000 10000
```

In this case, with ofwSVM, the user should set `mtry= 5` and `nsvm= 30000` for the learning step if `nstable=25`.

For both classifiers, we strongly advise to choose the smallest `mtry` that gives the more stable results. Our experience shows that for ofwCART, `mtry` will be rather small (5 to 15), as the trees are aggregated. For ofwSVM, `mtry` will usually be larger (> 15). In both cases, `mtry` should not be greater than `nstable`, and, therefore, `mtryTest` $\leq$ `nstable`.

Table 1 illustrates the tuned parameters for several public data sets that were tested in Lê Cao et al. (2007b) and Lê Cao et al. (2007a) for the weighted (ofwCART, ofwSVM) and non weighted (w-ofwCART, w-ofwSVM) versions of OFW.

3.3 Variable selection and visualization plots

Once the parameters `mtry`, and `ntree` for ofwCART, have been chosen, the variable selection step (**step 3** Figure 1) can be performed, preferably on the whole data set if

Table 1: Values of the size of the subset ω.

	#genes	#classes	#obs.	ofwCART	w-ofwCART	ofwSVM	w-ofwSVM
Lymphoma	4026	3	62	5^1	10^1	5	5
Leukemia	3000	3	72	5^1	5^1	15	10
SRBCT	2308	4	63	5^1	10^1	20	20
Brain	1963	5	42	5^1	25^1	10	10
Follicle	1564	3	42	10^2	10^2	25	25

The number of trees aggregated is ^{1}`ntree`= 150 and ^{2}`ntree`= 100.

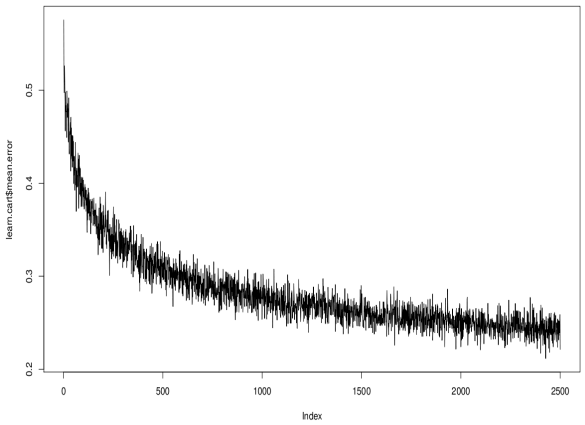


Figure 2: Internal mean error in ofwCART.

the sample size is too small, i.e., if n is roughly less than 80, or if the number of observations per class is too small. We advise to use the total number of iterations `nforest` or `nsvm`, rather than the `nstable` early stopping criterion to halt the algorithm, as suggested in Section 3.2.

The classifier to be applied has to be specified by the user. Here is the command for the variable selection step (**step 3**) for ofwCART and ofwSVM.

```
R> learn.cart <- ofw(srbct, as.factor(class), type="CART", ntree=150,
+ nforest=2500, mtry=5)
R> learn.svm <- ofw(srbct, as.factor(class), type="SVM", nsvm=30000, mtry=5)
```

In the case of ofwCART, the evolution of the internal mean error rate $\bar{\epsilon}_i = \frac{1}{\text{ntree}} \sum_{b=1}^{\text{ntree}} \epsilon_i^b$ can be plotted for each iteration i, as shown in Figure 2, for $i = 1, \ldots,$ `nforest`:

```
R> plot(learn.cart$mean.error, type="l")
```

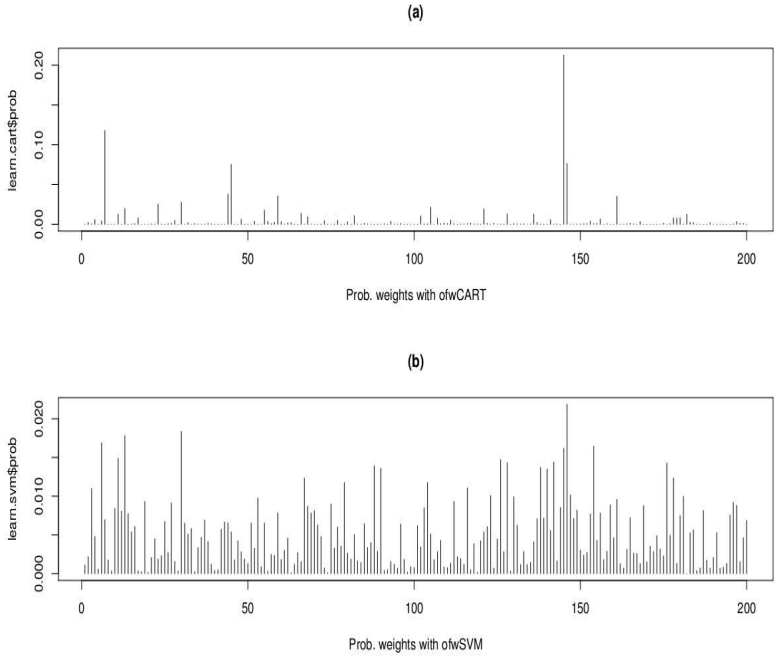


Figure 3: Variable weights that are computed with ofwCART **(a)** and ofwSVM **(b)**.

The monotonic decreasing trend of $\bar{\epsilon}_i$ indicates if the parameters have been tuned correctly and thus if ofwCART converges. In the case of ofwSVM, the SVM are not aggregated, and the error variance is consequently very large: no decreasing trend can be observed and ϵ_i is not provided. Note that the internal error $\bar{\epsilon}_i$ does not evaluate the performance of OFW (see below Section 3.4) and is simply a way to assess the quality of the tuning.

One can also visualize the probability weights $\mathbb{P}_{\texttt{nforest}}$ or $\mathbb{P}_{\texttt{nsvm}}$ for each variable (Figures 3**(a)** and **(b)**):

```
R> plot(learn.cart$prob, type="h")
R> plot(learn.svm$prob, type="h")
```

The selected variables can then be extracted by sorting the heaviest weights in $\mathbb{P}$, here for example for the 10 most discriminative variables:

```
R> names(learn.cart$list[1:10])
```

As $\mathbb{P}$ is a weight probability, the more numerous the variables, the smallest the weights on the variables. Hence, these weights are a qualitative rather than a quantitative importance measure of the variables, and the choice of a threshold is not advised. The

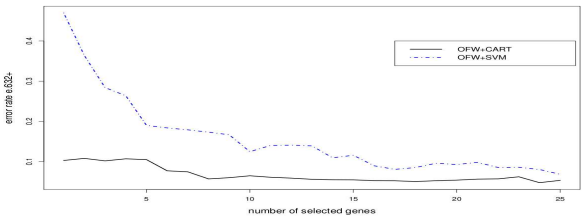


Figure 4: e.632+ error rate of ofwCART and ofwSVM.

different computations of the approximated gradient in ofwSVM (d_i) and in ofwCART (D_i), where $D_i >> d_i$, actually lead to an important number of weights in $\mathbb{P}$ close to zero in ofwCART. Remark that some of the very discriminative variables get important weights in both methods, but usually, as the classifiers SVM and CART are differently constructed, the resulting variable selections will not be the same (see Lê Cao et al. 2007b,a).

3.4 Evaluation step

Method and implementation

To assess the performance of the variable selection performed by OFW (**step 2** in Figure 1), we propose to perform the e.632+ bootstrap error estimate from Efron and Tibshirani (1997) that is adequate for small sample size data sets (Ambroise and McLachlan, 2002). Note that e.632+ does not dictate the optimal number of features to select. The error rate estimates that are computed with respect to the number of selected variables are only a way to compare the performances of different variable selection methods. **Step 2** consists in two functions called `learn` and `evaluate`. The `learn` function simply learns OFW on a fixed number of bootstrap samples (`Bsample`) with the same tuned parameters defined in **step 1**. The `evaluate` function that was inspired from the `ipred` package, computes and outputs the e.632+ error rate.

The learn and evaluate functions

```
R> learn.error.cart <- learn(srbct, as.factor(class), type="CART", ntree=150
+ nforest=2500, mtry=5, Bsample=10, do.trace=100, nstable=25)
R> learn.error.svm <- learn(srbct, as.factor(class), type="SVM", nsvm=30000,
mtry=5, Bsample=10, do.trace=2000, nstable=25)
```

As the evaluation will be performed for a small selection size, we strongly advise to reduce the number of total iterations, using for example the early stopping criterion. In the literature, `Bsample` often equals to 10-50. On a 1.6 GHz 960 Mo RAM AMD

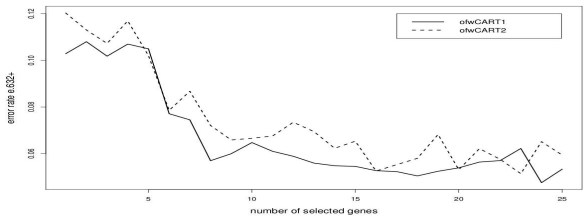


Figure 5: e.632+ error rate of ofwCART with e.632 not weighted (ofwCART1) and weighted (ofwCART2) .

Turion 64 X2 PC, the learning step of one bootstrap sample on a typical microarray data set ($p \simeq 5000$ and $n \simeq 50$) can take approximatively 2.5 hours. Hence, depending on the chosen value of `Bsample`, this evaluation step might be time consuming (see Section 4) and one can rather choose to perform parallel computing using the `Rmpi` package (see supplemental data).

If the SVM classifier is applied, each SVM is evaluated with the heaviest variables in $\mathbb{P}^b_{\texttt{nsvm}}$, which is learnt in the `learn` function, $b = 1, \ldots,$ `Bsample`. If the CART classifier is applied, the `evaluate` function aggregates `ntreeTest` trees. Each tree is constructed on a small variable subset that is randomly selected from the heaviest variables in $\mathbb{P}^b_{\texttt{nforest}}$, to avoid a too optimistic evaluation (see Lê Cao et al. 2007a). Both functions evaluate the variable selection of size `maxvar`:

```
R> eval.error.cart <- evaluate(learn.error.cart, ntreeTest=100, maxvar=25)
R> eval.error.svm <- evaluate(learn.error.svm, maxvar=25)
```

The `evalCARTparallel` function has also been implemented for parallel computing (refer to supplemental data). The aim of the `evaluate` function is to compare the performance of several algorithms (e.g., ofwCART and ofwSVM):

```
R> matplot(cbind(eval.error.cart$error, eval.error.svm$error), xlab="number
+ selected genes", ylab="error rate e.632+", type="l", col=c(1,4), lty=c(1,4
+ lwd=2, cex.lab= 1.3)
R> legend(18,0.40, c("ofwCART", "ofwSVM"), col=c(1,4), lty=c(1,4), cex=1.2,
+ lwd=2)
```

Figure 4 displays the e.632+ bootstrap error rate of the selections resulting from either ofwCART or ofwSVM with respect to the number of selected genes. In this example, where we compare the non-weighted versions of OFW, ofwCART seems to perform the best.

3.5 Further analysis

Comparing the weighted and non weighted versions of OFW

The weighting procedure presented in Section 3.4 has also been included in the error evaluation function `evaluate`. To compare the two approaches weighted (OFW) and non weighted (wOFW), we strongly advise to launch the `evaluate` function with the argument `weight=T` in *both* cases to evaluate if the minority classes were misclassified or not. Otherwise, the e.632+ bootstrap error rate will always be lower for OFW than wOFW. This is illustrated in Figure 5 where the same gene selection resulting from ofwCART is evaluated either with the non-weighted version of e.632+ (ofwCART1) or with the weighted version of e.632+ (ofwCART2). Even though the same two gene selections are evaluated, the error rate is lower in ofwCART1 as this overall error rate only takes into account the microarrays that are rightly classified in the majoritary classes. In ofwCART2 where misclassified minoritary classes are taken into account, the error rate is consequently higher:

```
R> learn.error.cart=learn(srbct, as.factor(class), type="CART, ntree=150,
+ nforest=3000, mtry=5, Bsample=10)
R> eval.error.cart1=evaluate(learn.error.cart, ntreeTest=100, maxvar=25)
R> eval.error.cart2=evaluate(learn.error.cart, ntreeTest=100, maxvar=25,
+ weight=T)
R> matplot(cbind(eval.error.cart1$error, eval.error.cart2$error), xlab=
+ "number of selected genes", ylab="error rate e.632+", type="l",
+ col=c(1,1), lty=c(1,2), lwd=2, cex.lab=1.3)
R> legend(18,0.12, c("ofwCART1", "ofwCART2"), col=c(1,1), lty=c(1,2),
+ cex=1.2, lwd=2 )
```

4 Computation time

Optimal Feature Weighting is a stochastic method that might be computationally time consuming if the variable dimension is very high. As the algorithm gets stabler for a large number of iterations, the variable selection step (**step 3**) might take 1-2 hours. Therefore, using parallel computing with the `Rmpi` package during the evaluation step (**step 2**) might be advisable. If the dimension is considerable, we strongly advise to pre-filter the data set so as to remove uninformative variables that slow down the computation.

In this paper, on a very small microarray data set (200 genes), the tuning step (**step 1**) took approximatively 20 min, the evaluation step (**step 2**) 1.5 hour and the variable selection step (**step 3**) 7 min.

5 Conclusion

We have implemented the stochastic algorithm 0FW to select discriminative features. Although we illustrated this method on microarray data, OFW can be applied on any continuous data set for classification and prediction purposes.

Wrapper methods usually require heavy computation, and so does OFW. Efforts have thus been made to reduce some of the computation time by implementing `C` functions when applying CART and by proposing parallel programming during the learning step.

With this package, we hope to provide the user a method with a strong theoretical background that is easy to apply and that can bring interesting results in a feature selection framework.

Availability and requirements

The `R` version $\geq$ 2.5.0 is needed to load the svm library `e1071`.

Acknowledgements

We are grateful to "Projet Calcul en MIdi-Pyrenées" (CALMIP) for the intensive computations, and the anonymous reviewers for their helpful comments on the manuscript.

References

Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction in tumour classification on basis of microarray gene expression data. Proc. Natl. Acad. Sci. USA, 99(1):6562–6566.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2):123–140.

Breiman, L. (2001). Random forests. Machine Learning, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA.

Chen, C., Liaw, A., and Breiman, L. (2004). Using random forest to learn imbalanced data. Technical Report 666, Dpt. of Statistics, University of Berkeley.

Chen, D., Hua, D., Reifman, J., and Cheng, X. (2003). Gene selection for multi-class prediction of microarray data. In CSB '03: Proceedings of the IEEE Computer Society Conference on Bioinformatics, page 492, Washington, DC, USA. IEEE Computer Society.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the e.632+ bootstrap method. Journal of American Statistical Association, 92:548–560.

Gadat, S. and Younes, L. (2007). A stochastic algorithm for feature selection in pattern recognition. J. Mach. Learn. Res., 8:509–547.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3):389–422.

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med, 7(6):673–679.

Lê Cao, K.-A., Bonnet, A., and Gadat, S. (2007a). Multiclass classification and gene selection with a stochastic algorithm. Technical report, Institut de Mathématiques, UMR CNRS 5219, University of Toulouse.

Lê Cao, K.-A., Gonçalves, O., Besse, P., and Gadat, S. (2007b). Selection of biologically relevant genes with a wrapper stochastic algorithm. Statistical Applications in Genetics and Molecular Biology, 6(:Iss. 1):Article 1.

Li, T., Zhang, C., and Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20(15):2429–2437.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. Rnews, 2/3(December):18–22.

Vapnik, V. N. (1999). The Nature of Statistical Learning Theory (Information Science and Statistics). Springer.

Weston, J., Elisseeff, A., Schölkopf, B., and Tipping, M. (2003). Use of the zero norm with linear models and kernel methods. J. Mach. Learn. Res., 3:1439–1461.

Yeung, K. and Burmgarner, R. (2003). Multi-class classification of microarray data with repeated measurements: application to cancer. Genome Biology, 4(R83).

Appendix

Parallel computing with ofwCART

```
## This is an example to perform the learning and the evaluation step of ofw
# with parallel computing
## A part of this code has been borrowed from Rmpi examples of the Acadia Ce
# for Mathematical Modelling and Computation
## http://ace.acadiau.ca/math/ACMMaC/Rmpi/examples.html

library("Rmpi")
library(rlecuyer)
library(e1071, lib.loc="MyR/Library")#if the library e1071 is locally instal
library(ofw, lib.loc="MyR/Library")  #if the library ofw is locally installe

mpi.spawn.Rslaves(nslaves=5)  #number of slaves to spawn, should be equal to
                              #where B = the number of bootstrap samples
mpi.setup.rngstream() #generates random numbers

.Last <- function(){
    if (is.loaded("mpi_initialize")){
        if (mpi.comm.size(1) > 0){
            print("Please use mpi.close.Rslaves() to close slaves.")
            mpi.close.Rslaves()
        }
        print("Please use mpi.quit() to quit R")
        .Call("mpi_finalize")
    }
}

##----------FUNCTION ----------------------------------------

##learn ofw on the bootstrap sample
##assume all the parameters
learn.ofw = function(){
#if both libraries are locally installed:
library(e1071, lib.loc="MyR/Library")
library(ofw, lib.loc="MyR/Library")

x=data[mat.train[,foldNumber],]
```

```
y=class[mat.train[,foldNumber]]
res=ofw(x=x, y=as.factor(y), type=type, ntree=ntree, nforest=nforest,
nsvm=nsvm, mtry=mtry, do.trace=do.trace, nstable=nstable
, weight=weight)
return(list(res$prob))
}

##----------MAIN ------------------------------------------

# We are  in the parent.
#read the data set here attached in the library ofw

data(srbct)
attach(srbct)
data=srbct
class=class

#define parameters and constants
B=5   #number of bootstrap samples
nvar=ncol(data)
nobs=nrow(data)
mtry=5
do.trace=FALSE
nstable=FALSE
weight=F

#parameters to learn and evaluate ofwCART:
type= "CART"
ntree=10
nforest=10

#during evaluation
maxvar = 10
ntreeTest = 15

#define the matrices
mat.train = matrix(nrow=nobs, ncol=B)
mat.P=  matrix(nrow=nvar, ncol=B)

#define the bootstrap samples
```

```
for(i in 1:B)
{
again=T
while(again){
train=sample(1:nobs, nobs, replace=T)
if(any(table(class[train])==0)) {again=T} else {again=F}
}
mat.train[,i]=train
}

nslaves= mpi.comm.size() -1

#send parameters and constants to slaves
mpi.bcast.Robj2slave(data)
mpi.bcast.Robj2slave(class)
mpi.bcast.Robj2slave(B)
mpi.bcast.Robj2slave(nvar)
mpi.bcast.Robj2slave(nobs)
mpi.bcast.Robj2slave(mtry)
mpi.bcast.Robj2slave(do.trace)
mpi.bcast.Robj2slave(nstable)
mpi.bcast.Robj2slave(weight)
mpi.bcast.Robj2slave(type)
mpi.bcast.Robj2slave(ntree)
mpi.bcast.Robj2slave(maxvar)
mpi.bcast.Robj2slave(ntreeTest)
mpi.bcast.Robj2slave(nforest)
mpi.bcast.Robj2slave(mat.train)
mpi.bcast.Robj2slave(mat.P)


#send functions to slaves
mpi.bcast.cmd(foldNumber <- mpi.comm.rank())
mpi.bcast.Robj2slave(learn.ofw)

#each slave learns ofwCART on each bootstrap sample
res.slaves = mpi.remote.exec(learn.ofw(), comm=1)

#get the results
for (i in 1:nslaves)
```

```
{
mat.P[,i]=res.slaves[[i]][[1]]
}

#once the probability has been learnt on each bootstrap sample,
#evaluate the selection with the function evaluateCARTparallel.R
res.eval=evaluateCARTparallel(x=data, y=as.factor(class), matTrain = mat.tra
matProb= mat.P, maxvar = maxvar, ntreeTest=ntreeTest, weight=weight)

#write output
write.table(mat.train, "mat.train.txt")
write.table(mat.P, "mat.P.txt")
write.table(res.eval$error, "error.txt")

detach(srbct)

mpi.close.Rslaves()
mpi.quit(save="yes")
```

6. Bilan et perspectives

Nous avons montré dans cette partie les difficultés rencontrées pour évaluer statistiquement la performance de la sélection, compte tenu du nombre restreint d'échantillons. C'est pourquoi nous avons jugé nécessaire de toujours nous placer à l'interface entre la statistique et la biologie, afin de pouvoir répondre à la question : "est-ce que l'approche que nous proposons est utile et répond correctement aux attentes des biologistes ?"

Dans ce travail, nous n'avons pas répondu à la question de la taille optimale de la sélection. De façon générale, nous avons vu que peu d'approches de types *wrapper* ou *embedded* s'intéressent à ce sujet difficile. De plus, tout dépend du contexte biologique. Est-il utile pour le biologiste de sélectionner une liste de gènes-marqueurs réduite, sans information redondante ? Ou, au contraire, de sélectionner une liste suffisamment importante pour identifier des gènes discriminants expliquant de manière globale l'expérience ? Il faut aussi tenir compte du fait que de nombreux gènes mesurés sur la puce sont encore inconnus et non annotés (au moins dans le cadre de la génétique animale). Par conséquent, lors de son travail d'interprétation, le biologiste a besoin d'une liste de gènes suffisamment importante pour qu'elle soit informative. Il parait pour le moment cohérent que ce soit le biologiste qui décide du nombre de gènes qu'il veut sélectionner, et non pas le statisticien.

Il n'est pas possible pour le moment d'appliquer OFW avec des données manquantes. Celles-ci pourraient être estimées avec les k plus proches voisins (k-NN), comme le proposent Dudoit *et al.* (2002), en sélectionnant des gènes avec des profils d'expression similaires à celui dont on veut estimer les valeurs manquantes. Une autre approche serait la décomposition en valeurs singulières (Valafar, 2002; Troyanskaya *et al.*, 2001) ; cependant elle semble moins robuste.

Par ailleurs il serait intéressant d'inclure dans OFW une stratégie boosting, afin de donner plus de poids à des échantillons difficiles (appartenant par exemple à des classes minoritaires). Il serait aussi intéressant de proposer à l'utilisateur plusieurs types de SVM pour traiter le cas multiclasse (*one-vs-rest*, ou le SVM multiclasse de Lee & Lee, 2003).

Enfin, il aurait été intéressant de comparer l'algorithme OFW à l'algorithme RELIEF de Kira & Rendell (1992) et à ses nombreuses variantes : (ReliefF, Kononenko, 1994, pour le multiclasse ; RReliefF, Robnik-Sikonja & Kononenko, 1997, dans le cas de la régression ; et, plus récemment, I-Relief, Sun & Li, 2006 qui éliminent les données

aberrantes et modifient la fonction définissant la marge). De la même façon que OFW, les algorithmes RELIEF adoptent une approche de pondération de variables, qui consiste à optimiser un problème convexe basé sur la maximisation d'une fonction de marge. Cette marge est définie par le classifieur 1-NN. Cet algorithme semble très efficace, car il ne nécessite pas une recherche exhaustive ou combinatoire comme les méthodes de type *wrapper*. Le principe repose sur une estimation itérative des poids sur les variables selon leur capacité à discriminer des observations appartenant à la même classe. A chaque itération, une observation $\mathbf{x}$ est aléatoirement choisie, ainsi que ses deux plus proches voisins, l'un de la même classe (*nearest hit*, NH) et l'autre d'une autre classe (*nearest miss*, NM). Le poids w_j de la jème variable est ensuite mis à jour :

$$\mathbf{w}_j = \mathbf{w}_j + |\mathbf{x}^j - NM^j(\mathbf{x})| - |\mathbf{x}^j - NH^j(\mathbf{x})|.$$

De son côté, OFW s'appuie sur une pondération adaptative des variables, en favorisant le tirage des variables les plus discriminantes et comportant de l'information complémentaire, selon une certaine probabilité qui est estimée itérativement dans l'algorithme. Ces deux approches pourraient donc donner des résultats différents mais néanmoins intéressants.

Pour finir, certains développements théoriques mériteraient d'être étudiés pour comprendre la convergence de OFW associé à un gradient "moyenné" qui permet d'accélérer la convergence lorsque CART est agrégé.

Deuxième partie

Sélection de variables pour l'intégration de données "omiques".

7. Contexte

7.1 Motivations

7.1.1 Les données omiques

Les avancées technologiques permettent maintenant de générer des données de grande dimension issues de plateformes différentes. Ainsi l'analyse de données transcriptomiques, protéomiques et métabolomiques sur un même système biologique est maintenant possible. Cette approche analytique globale, encore appelée biologie intégrative (*systems biology*), a pour but d'étudier un ou des organismes en intégrant des données de sources multiples, et ainsi de mieux appréhender les différents processus cellulaires très complexes inhérents au système.
Commençons par quelques définitions très générale concernant ces données (figure 7.1).
Le *métabolome* est constitué de petites molécules, les métabolites, présents dans un organisme (cellules, tissus, organes...), qui peuvent être considérées comme l'expression ultime des gènes en réponse à des changements environnementaux. En effet les métabolites sont régulés (entre autres) par l'activité d'enzymes, qui sont elles mêmes dépendantes du niveau d'expression des gènes. Contrairement aux données transcriptomiques ou protéomiques, les profils métaboliques permettent de donner un aperçu de la physiologie d'une cellule.
Les protéines synthétisent ou dégradent les métabolites, et leur ensemble constitue le *protéome* d'un organisme ou système. Là encore les données sont complexes puisque le protéome varie d'une cellule à l'autre et subit des changements constants dûs à des interactions biochimiques avec le génome et l'environnement (à l'instar du génome, qui reste relativement constant). La complexité s'accroit aussi du fait que plusieurs protéines peuvent dériver d'un seul gène.
Le *transcriptome* est l'ensemble des ARN produits à partir du génome à un instant donné. Toutes les cellules d'un organisme pluricellulaire ont le même patrimoine génétique, leur *génome*, mais elles diffèrent dans leurs fonctions et leurs formes car elles n'utilisent pas la même combinaison des informations contenues dans le génome, c'est-à-dire qu'elles n'expriment pas le même panel de gènes. Les cellules sont capables de s'adapter à une modification de leur environnement en modifiant l'expression de leurs gènes. Une modulation de l'expression des gènes se traduit par des modifications de l'abondance des ARN qui sont produits à partir des gènes, puis par une modification de l'abondance des protéines correspondantes et enfin par des modifications des activités de ces protéines.
En plus du schéma proposé figure 7.1, d'autres niveaux de régulation pourraient être

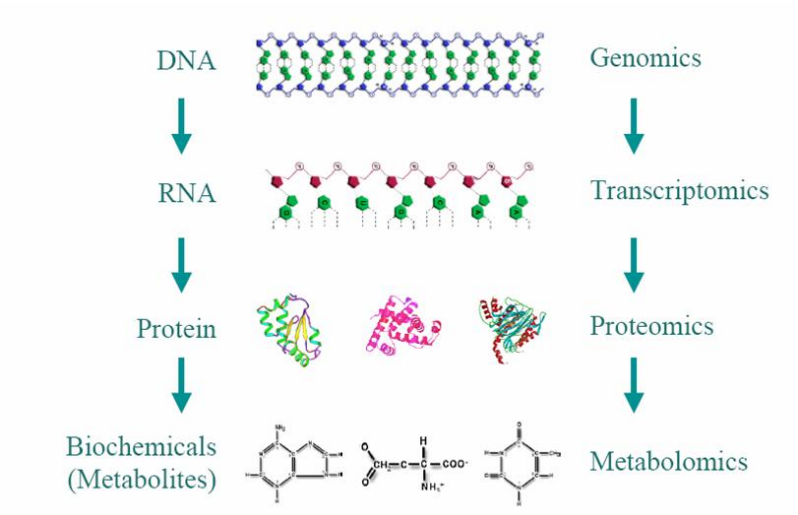


Fig. 7.1: Les données "omiques" et les différents niveaux fonctionnels.

envisagés, comme la régulation de la traduction des ARN messagers en protéines ou la régulation de l'activité des protéines par des modifications post-traductionnelles, des interactions avec d'autres protéines, des modification de leur localisation subcellulaire ou par la disponibilité des substrats, produits ou cofacteurs des réactions enzymatiques. Nous nous intéressons dans ce travail qu'aux trois niveaux fonctionnels : transcriptome, métabolome et protéome.

Malgré le séquençage complet de plusieurs espèces (Saccharomyces cerevisiae, Escherichia coli, Arabidopsis thaliana, Oryza sativa, Drosophila melanogaster, Homo sapiens...), de nombreux gènes, fonctions de gènes, ainsi que les réseaux gènes-métabolites ou gènes-protéines, restent inconnus. Si l'identification de ces fonctions peut être faite de façon ciblée sur quelques métabolites (protéines) en fonction du niveau d'expression des gènes, une approche plus globale est nécessaire pour décrire tout un processus cellulaire. Cependant l'analyse de mesures de gènes (connus ou inconnus), de métabolites et/ou de protéines non ciblés lors de conditions expérimentales identiques reste difficile sans l'aide de méthodes statistiques adéquates (Hirai *et al.*, 2004).

Les objectifs d'une telle analyse intégrative sont les suivants (Steinfath *et al.*, 2008) :

1. comprendre les interactions entre variables omiques de même type ;
2. relier ces variables omiques de même type au phénotype (ensemble de caractéristiques observables caractérisant un être vivant) ;
3. identifier (ou préciser) la fonction de gènes inconnus (peu connus) ;
4. comprendre les relations entre des variables omiques de types différents.

Ces objectifs rendent l'analyse statistique peu aisée compte tenu du nombre limité d'échantillons et du grand nombre de variables mesurées.
Dans cette partie, nous nous intéressons tout particulièrement au dernier point, dans le cas de deux types de données omiques mesurées.

7.1.2 Intégration

L'intégration de données biologiques de types omiques est relativement récente dans la littérature, avec des premières études apparues dans Hirai *et al.* (2004, 2005) pour l'intégration de données métabolites et d'expression de gènes chez *Arabidopsis thaliana*. Les méthodes appliquées pour résoudre ces problèmes sont très souvent des méthodes classiques de projection multivariées, puisqu'elles permettent de projeter les données dans des espaces de dimension plus petite. Le biologiste peut alors plus facilement interpréter les résultats grâce à des représentations graphiques résumant l'information.

7.1.3 Méthodes d'analyse multivariées

Il existe de nombreuses méthodes dites d'analyse multivariées : Analyse en Composantes Principales, Analyse des Corrélations Canoniques, Analyse Factorielle Discriminante, Analyse en Composantes Indépendantes... Le lecteur pourra se référer à Mardia *et al.* (1979) pour une description de ces approches, ainsi qu'aux revues de Baccini *et al.* (2005); Steinfath *et al.* (2008) pour une étude comparative de certaines de ces méthodes sur des données biologiques.
Dans le cadre de ce travail, nous nous sommes principalement focalisés sur trois approches.
La première, en tant que méthode de référence, est l'Analyse en Composantes Principales (PCA), dans le cadre d'*analyse à un seul tableau*. Cette méthode clef permet de définir la plupart des méthodes dites "factorielles"' dont la solution est obtenue soit par diagonalisation d'une matrice carrée (matrice de covariance ou de corrélation), soit par la décomposition en valeurs singulières (SVD) de la matrice centrée des données initiales. C'est de loin la méthode la plus utilisée dans le cas d'une analyse à un tableau. Dans le cadre de données de très grande dimension, nous nous sommes inspirés des variantes dites *sparse* PCA pour répondre à notre objectif.
Les deux autres approches sur lesquelles nous nous sommes aussi penchés s'inscrivent dans l'*analyse de deux tableaux*, afin de mettre en relation des données omiques de deux types différents, mais mesurées sur les mêmes individus ou échantillons. Il s'agit de l'-Analyse des Corrélations Canoniques (CCA) et de la régression Partial Least Squares (PLS).

7.1.4 Pallier le problème de la grande dimension

Dans le cas particulier de données de grande dimension, ces trois approches sont cependant limitées par le nombre de variables, soit par manque de lisibilité (PCA, PLS), soit à cause de problèmes numériques (CCA). Une des solutions proposées pour

pallier ce problème est d'adapter des idées maintenant bien connues dans le cadre de la *régression* qui sont basées sur la prise en compte de pénalités. Ainsi, une pénalité de type norme l_2, conduit à la régression Ridge (Hoerl & Kennard, 1984), qui permet de régulariser des matrices singulières non inversibles. Ce principe a été appliqué à la comparaison de deux tableaux avec la CCA par Vinod (1976). Des pénalités de type norme l_1, encore appelée *Lasso* (Tibshirani, 1996), ou bien de type norme l_0 ont aussi été proposées pour permettre la sélection de variables. Notons encore l'algorithme de sélection nommé LARS ("Least Angle Regression", Efron *et al.*, 2004), ainsi que l'approche appelée *elastic net*, résolue de façon efficace via LARS, et qui consiste à combiner à la fois des pénalités de types l_1 et l_2 (Zou & Hastie, 2005).
Ce deuxième type de pénalisations, appliqués aux vecteurs de poids des variables (encore appelés *loadings*) issus de la PCA, la CCA ou la PLS, permettent de faire de la sélection de variables sur un tableau (cas de la PCA) ou deux tableaux (CCA, PLS). Dans le cas de deux tableaux, la sélection faite sur chaque paquet de variables permet de mettre en relation les variables les plus importantes (mais de types différents) dans le contexte de la biologie intégrative.

7.1.5 Objectifs

Le but de ce chapitre est de présenter certaines régularisations/pénalisations qui ont été appliquées sur les méthodes qui nous intéressent : PCA, CCA et PLS, afin d'introduire l'approche *sparse PLS* que nous avons développée dans ce travail. Par conséquent nous nous plaçons directement dans le cadre d'un double objectif de "sélection et intégration" sans expliciter les méthodes de pénalisation en régression (l_0, Lasso, Ridge ou Elastic Net) qui sont largement explicitées dans la littérature et référencées dans ces travaux. Les méthodes PCA et CCA seront brièvement présentées et seule l'approche PLS sera développée plus en détail afin de bien expliciter la suite des travaux.

Il est important de noter que, dans cette partie, la problématique de sélection de variables s'inscrit dans un cas d'*intégration* de données de type différents, et non pas comme vu dans la partie 2 dans un contexte de discrimination ou classification supervisée des observations ou échantillons. Le but des travaux que nous introduisons ici est de réaliser en une seule procédure la sélection des variables omiques les plus pertinentes en vue de l'intégration de deux ensembles de données. D'autre part, à l'aide de graphiques appropriés, ce genre d'approche permet à la fois de faciliter l'interprétation des résultats pour le biologiste, et de mettre en valeur certaines relations entre variables de types différents. Si les variables que nous sélectionnons sont encore inconnues, ou si les relations que nous soulignons entre ces deux groupes de variables sont encore inconnues, cela oriente le biologiste vers de nouvelles expériences pour vérifier ces hypothèses.

7.1.6 Notations

Dans cette partie, nous utilisons les notations suivantes : les données sont résumées dans deux matrices X de taille $n \times p$ et Y de taille $n \times q$, où les vecteurs x^j et y^k représentent les mesures des variables j et k de type X ou Y pour chaque échantillon. Nous nous plaçons dans le cadre de données de grandes dimensions, où $p + q >> n$.

7.2 Méthode d'analyse à un tableau : PCA

L'Analyse en Composantes Principales (Jolliffe, 2002) est la méthode de projection multivariée la plus connue. En général elle ne s'applique que sur un seul des deux tableaux, X ou Y. La PCA est utilisée comme un outil préliminaire pour visualiser de façon rapide si les échantillons biolologiques peuvent être séparés au niveau de l'expression des variables suivant les conditions biologiques mesurées lors de l'expérience (Steinfath *et al.*, 2008).
Rappelons que le but de la PCA est de trouver des combinaisons linéaires des variables initiales appelées *composantes principales*, qui maximisent la variance du jeu de données. Les composantes principales sont orthogonales entre elles et représentent de nouvelles variables artificielles non correlées. Ainsi, nous recherchons les vecteurs unitaires $v_1 \dots v_H$ tels que

$$\arg \max_{v_h' v_h = 1} var(X v_h) \tag{7.1}$$

où les v_h, $h = 1 \dots H$, $H < p$ ou q, sont les vecteurs appelés facteurs principaux ou *loadings* et les composantes principales associées sont les Xv_h. La plus grande partie de la variance est, par construction, expliquée par les premières composantes principales H. Notons une propriété très utile des facteurs principaux qui est la correspondance directe entre leur coordonnées et l'importance des variables dans le modèle, dans le cas de variables homogènes ou réduites.
Deux façons de résoudre le problème (7.1) sont possibles :
-Par la résolution des problèmes de vecteurs propres et valeurs propres de la matrice des covariances ou de celle des corrélations (Jolliffe, 2002). Dans ce cas les valeurs propres sont égales à la variance expliquée par chaque composante principale et les vecteurs propres associés sont les facteurs principaux.
-En utilisant l'algorithme NIPALS (Nonlinear estimation by Iterative Partial Least Squares, Wold, 1966) qui est à la base de la régression PLS. Cet algorithme résout la recherche des valeurs et vecteurs propres par l'algorithme de la puissance itérée. Nous reviendrons sur cette résolution particulière du problème dans la section 7.3.2.
Une limite de la PCA lorsque l'on dispose de données de grande dimension est l'interprétabilité des résultats lorsque le nombre de variables devient trop grand. Plusieurs méthodes de PCA "parcimonieuses", ou *sparse PCA*, ont ainsi été proposées dans la littérature pour pallier ce problème. Chacune de ces approches présente une résolution différente du problème pour faire de la sélection de variables sur chaque dimension de la PCA.

Une des premières approches de sparse PCA fut le *simple thresholding* ou simple seuillage, qui consiste, de manière empirique, à annuler les coefficients dont les valeurs absolues sont inférieures a un seuil donné (Cadima & Jolliffe, 1995). Par la suite, d'autres approches accompagnées de justifications mathématiques (et surtout numériques) furent proposées. Les principales limites de toutes ces approches sparse PCA sont la perte d'orthogonalité des composantes principales et, surtout, le fait qu'une part de variance expliquée (que l'on espère petite) est perdue par rapport à une PCA classique. Ainsi, lorsqu'une pénalité trop forte est appliquée sur chaque facteur principal (sélection trop forte), les composantes principales sont très correlées et l'interprétation graphique devient très difficile. Et, plus la pénalité est forte, plus le pourcentage de variance expliquée pour chaque composante PCA devient faible. Ceci pousse l'utilisateur à choisir un grand nombre de composantes principales (ce qui rend l'interprétabilité des résultats bien moins aisée), ou bien à pénaliser moins et à obtenir un modèle peu parcimonieux.

Nous présentons brièvement ici quatre approches sparse PCA différentes qui ont été développées ces dernières années.

Scotlass. Cette approche proposée par Jolliffe *et al.* (2003) puis Trendafilov & Jolliffe (2006), consiste à pénaliser les facteurs principaux de la PCA v_h au moyen d'une pénalité lasso. Le problème à résoudre est le suivant :

$$\arg\max_{v_h} v_h'(X'X)v_h,$$

sous les contraintes habituelles de la PCA $v_h'v_h = 1$ et $v_h'v_k = 0$ pour $k < h$, avec $h \geq 2$. La condition lasso est ajoutée : $\sum_{j=1}^{p} v_{h,j} \leq t$, pour $j = 1 \dots p$, avec t tel que $t \leq \sqrt{p}$.
La résolution de ce problème d'optimisation avec contrainte se fait grâce à une approche de gradient projeté et nécessite l'usage d'approximations pour inclure la pénalité lasso. Les principales critiques que l'on peut faire à cette méthode, en plus de ce qui a été dit plus haut, sont le réglage du paramètre t, peu évoqué, et le temps de calcul important, le problème d'optimisation n'étant pas convexe.
Notons que les auteurs ont aussi proposé la méthode "DaLASS" (Trendafilov & Jolliffe, 2007) basée sur le même principe, mais pour l'analyse discriminante linéaire.

sparse PCA Elastic Net. Une autre approche proposée par Zou *et al.* (2006) est basée sur Elastic Net. Les auteurs reformulent la PCA dans un cadre de régression pour résoudre le problème

$$\arg\min_{\alpha,\beta} \sum_{i=1}^{n} |X_i - \alpha\beta' X_i|^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2 + \sum_{j=1}^{K} \lambda_{1,j}|\beta_j|$$

$$\text{sous la contrainte} \quad \alpha'\alpha = I_K$$

où α et β sont des matrices de taille $p \times K$, K est le nombre de composantes principales choisies, $\{\lambda_{1,j}\}$ et $\{\lambda_2\}$ sont les paramètres de pénalisation et X_i correspond au vecteur

ligne de X. Ici, les paramètres $\lambda_{1,j}$ permettent de pénaliser de façon différente chaque composante principale. Le problème se résout de façon itérative en calculant d'abord la matrice β puis $\alpha = UV'$ où U et V sont issus de la décomposition SVD de $\beta = UDV'$. Lorsque $p >> n$ les auteurs proposent une simplification de l'algorithme afin d'accélérer le temps de calcul avec une règle de seuillage (*soft thresholding*) très similaire à une penalité lasso.

Direct Sparse PCA (DSPCA). Aspremont *et al.* (2007) proposent d'incorporer directement le critère de parcimonie dans la formulation de la PCA, puis d'introduire des contraintes de relaxation au problème d'optimisation. Ceci permet de rendre le problème convexe. On obtient alors un problème de programmation semi-défini (SDP : *semi definite programming*) qui peut se résoudre avec des méthodes génériques de SDP ou bien, si $p >> n$, avec des méthodes de résolution de problèmes dits *saddle-point*. Les auteurs ne développent pourtant pas cette méthode de résolution, ce qui semble rendre cette approche sparse PCA peu efficace.
Le problème à résoudre est le suivant :

$$\max v'X'Xv \quad \text{sous les contraintes} \quad ||v||_0 \leq \kappa, ||v||^2 = 1 \tag{7.2}$$

avec κ fixé contrôlant la "sparsité". Ce problème étant non convexe, la contrainte $||v||_0 \leq \kappa$ est relaxée dans un problème semi-défini (voir Aspremont *et al.*, 2007). Un fois le premier vecteur sparse v_1 résolu, la matrice $X'X$ est mise à jour dans l'algorithme de déflation pour obtenir

$$(X'X)_2 = X'X - (v_1'X'Xv_1)v_1v_1',$$

ce qui permet de calculer à chaque déflation la composante principale suivante v_2, etc. Les auteurs proposent d'arrêter les itérations lorsque les éléments de la matrice $(X'X)_i$ deviennent proches du niveau du bruit $\rho^\star$ qu'ils définissent dans leur article. Les principaux désavantages de cette approche semblent être la complexité de l'algorithme de résolution, qui est de $\mathcal{O}(p^4\sqrt{log(p)})$ ainsi que le réglage du paramètre κ sur chaque dimension, qui n'est pas explicité.

Sparse SVD. Shen & Huang (2007) proposent de pénaliser les composantes de la PCA dans un cadre analogue à celui de la régression, en utilisant les approximations de rangs inférieurs de matrices grâce à une décomposition SVD.
Rappelons que la décomposition en valeurs singulières (SVD) de n'importe quelle matrice X de rang r s'écrit $X = UDV'$ et que pour un rang $l < r$, $X^{(l)} = \sum_{k=1}^{l} d_k u_k v_k'$ est la meilleure approximation de X de rang l, satisfaisant la minimisation du carré de la norme de Frobenius

$$\min_{X^*} ||X - X^*||_F^2 = tr(X - X^*)(X - X^*)'.$$

Dans le cas de la PCA, les auteurs s'intéressent à l'approximation de rang 1 de X s'écrivant $\tilde{u}\tilde{v}'$ et au problème d'optimisation :

$$\min_{\tilde{u},\tilde{v}} ||X - \tilde{u}\tilde{v}'||_F^2. \tag{7.3}$$

Les propriétés de la SVD avec les estimations de matrices de rang inférieur donnent directement la solution de (7.3) : $\tilde{u} = u_1$ et $\tilde{v} = d_1 v_1$. Le but de Shen & Huang (2007) est alors de relier l'équation (7.3) à un problème de régression des moindres carrés, afin d'introduire une pénalité (de type lasso par exemple) sur le facteur principal $\tilde{v}$.
Pour un vecteur $\tilde{u}$ fixé, le $\tilde{v}$ optimal devient en effet le vecteur des coefficients de régression des colonnes de X sur $\tilde{u}$, ce qui permet d'introduire une pénalité sur $\tilde{v}$. Le problème (7.3) devient alors :

$$\min_{\tilde{u},\tilde{v}} ||X - \tilde{u}\tilde{v}'||_F^2 + \lambda \sum_{j=1}^{p} |\tilde{v}_j| \tag{7.4}$$

dans le cas d'une pénalisation lasso. D'autres types de pénalité sont aussi proposés, comme le "hard thresholding" (Donoho & Johnstone, 1992) et le "smoothly clipped absolute deviation" (Fan & Li, 2001). Le problème (7.4) se résout de façon itérative très efficacement. Le réglage du paramètre de pénalisation λ se fait par validation croisée ou bien, de façon plus heuristique, en tenant compte du degré de sparsité et de la variance expliquée sur chaque composante.

Remarques

Les précédents auteurs ont tous appliqué leur approche sur le jeu de données *pitprops* connu pour sa difficulté dans l'interprétation des composantes principales (Jeffers, 1967). Pitprops ne contient cependant que 13 variables pour 180 observations. Pour une PCA classique, 6 composantes principales permettent d'expliquer 87% de la variance totale contre 78.2 % pour SCoTLASS, 75.8% pour Elastic Net sPCA et 84.7% pour SVD-sparse PCA. Le tableau 7.1 souligne les différences entre chaque approche et montre que, sans interprétation, la comparaison de ces approches reste d'un intérêt limité.
Zou *et al.* (2006) ainsi que Shen & Huang (2007) ont en plus fait l'effort d'appliquer leurs approches sur des données microarrays de grande dimension : Ramaswamy (Ramaswamy *et al.*, 2001) avec 16 063 gènes et NCI60 avec 2267 gènes. Des analyses étudiant la pertinence biologique de ces approches sparse PCA semblent maintenant nécessaires pour convaincre le biologiste de l'utilité de ces méthodes.

Dans le même état d'esprit que Aspremont *et al.* (2007), Aspremont *et al.* (2008) développent une méthode gourmande (*greedy*) de résolution de

$$\max_{||v||\leq 1} v'X'Xv - \kappa||v||_0$$

en se concentrant cette fois sur la pénalisation l_0 pour un coût $\mathcal{O}(p^3)$, puis reformulent le problème sparse PCA de façon convexe grâce à des contraintes de relaxation. Plusieurs variantes sparse PCA sont comparées, ainsi qu'une application sur les données Colon (Alon *et al.*, 1999) et Lymphoma (Alizadeh *et al.*, 2000). Un résultat intéressant est à noter : pour les données Colon, sur les 20 premiers gènes sélectionnés sur la composante 2 de la sparse PCA, 6 se retrouvent parmi les 10 premiers sélectionnés par le

	PCA			SCoTLASS			Elastic Net			SVD		
Variables	pc1	pc2	pc3	pc1	pc2	pc3	pc1	pc2	pc	pc1	pc2	pc3
topdiam	-0.404	0.218	-0.207	0.546	0.047	-0.087	-0.477	0.000	0.000	-0.449	0.000	0.000
length	-0.406	0.186	-0.235	0.568	0.000	-0.076	-0.476	0.000	0.000	-0.460	0.000	0.000
moist	-0.124	0.541	0.141	0.000	0.641	-0.187	0.000	0.785	0.000	0.000	-0.707	0.000
testsg	-0.173	0.456	0.352	0.000	0.641	0.000	0.000	0.620	0.000	0.000	-0.707	0.000
ovensg	-0.057	-0.170	0.481	0.000	0.000	0.457	0.177	0.000	0.640	0.000	0.000	0.550
ringtop	-0.284	-0.014	0.475	0.000	0.356	0.348	0.000	0.000	0.589	-0.199	0.000	0.546
ringbut	-0.400	-0.190	0.253	0.279	0.000	0.325	-0.250	0.000	0.492	0.399	0.000	0.366
bowmax	-0.294	-0.189	-0.243	0.132	-0.007	0.000	-0.344	-0.021	0.000	-0.279	0.000	0.000
bowdist	-0.357	0.017	-0.208	0.376	0.000	0.000	-0.416	0.000	0.000	-0.380	0.000	0.000
whorls	-0.379	-0.248	-0.119	0.376	-0.065	0.000	-0.400	0.000	0.000	-0.407	0.000	0.000
clear	0.011	0.205	-0.070	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
knots	0.115	0.343	0.092	0.000	0.206	0.000	0.000	0.013	0.000	0.000	0.000	0.000
diaknot	0.113	0.309	-0.326	0.000	0.000	-0.718	0.000	0.000	-0.015	0.000	0.000	-0.515
nb. var. sél.	13	13	13	6	7	7	7	4	4	7	2	4
cum var. (%)	32.4	50.7	65.1	27.2	42.5	56.9	28.0	42.0	55.3	30.6	45.0	59.0

Tab. 7.1: Comparaison des 3 premières composantes PCA sur les données pitprops pour les méthodes PCA, ScotLASS, Elastic Net sparse PCA et sparse PCA-SVD.

logiciel RankGene (Su *et al.*, 2003), bien que les auteurs ne spécifient pas la méthode utilisée (RankGene est un logiciel permettant d'utiliser au choix soit la méthode filtre test de Student, soit une méthode de discrimination par SVM dans un cas binaire). L'application d'une telle méthode semble cependant assez complexe en pratique.

7.3 Méthodes d'analyse à deux tableaux

7.3.1 CCA

Un outil adapté à l'analyse multivariée est l'Analyse Canonique des Corrélations (Hotelling, 1936) pour analyser, de façon symétrique, les relations entre deux tableaux de données. Ici, le but est de calculer des combinaisons linéaires des variables initiales des deux groupes, pour dans ce cas maximiser la corrélation :

$$\arg \max_{u_h' u_h = 1, v_h' v_h = 1} cor(X u_h, Y v_h) \qquad h = 1 \dots H.$$

On appellera ici le couple ($\zeta_h = X u_h$, $\eta_h = Y v_h$) les *variables canoniques*, et (u_h, v_h) les *facteurs canoniques* associés. Il est important de noter qu'ici, à l'instar de la PCA et de la PLS présentée plus loin (cf. section 7.3.2), les facteurs canoniques ne sont pas directement interprétables pour identifier l'importance des variables dans la mise en relation de X et Y (Gittins, 1985; Tenenhaus, 1998). En revanche on peut représenter les variables sur des cercles de corrélation, en projetant X et Y sur les espaces engendrés par les ζ_h et η_h, $h = 1 \dots H$. Les coordonnées obtenues étant les coefficients de corrélation entre les variables initiales et les variables canoniques (Tenenhaus, 1998; Saporta, 2006), on peut ainsi identifier les corrélations entre les deux groupes de variables. Une revue plus détaillée des représentations graphiques de la CCA se trouve dans González (2008, chap.3).
Le problème de la CCA se résout en calculant les vecteurs propres et valeurs propres du produit des projecteurs orthogonaux P_X et P_Y engendrés par les ζ_h et η_h respectivement, avec $P_X = X(X'X)^{-1}X'$ et $P_Y = Y(Y'Y)^{-1}Y'$. Sous cette forme, on comprend tout de suite la limitation de la méthode, puisque lorsque $p >> n$ et $q >> n$, le calcul des matrices inverses devient impossible. Il existe une autre méthode numérique pour résoudre ce problème, par décomposition SVD du produit des bases unitaires des espaces engendrés par X et Y (Bjorck & Golub, 1973), mais le calcul de ces bases est encore une fois limité lorsque les matrices sont mal conditionnées. On pourra se référer à González (2008) pour aborder les formulations d'un point de vue algébrique ou géométrique de la CCA.

On peut noter aussi l'approche *kernel-CCA* qui met en œuvre des méthodes à noyau (Scholkopf & Smola, 2001) pour plonger les données dans un espace de plus grande dimension (*feature space*), avant d'appliquer une CCA. Ici encore les problèmes rencontrés dans le cas $p + q >> n$ sont l'inversion des matrices kernel qui tendent alors à faire du surapprentissage et identifier des corrélations non pertinentes. Les solutions proposées sont soit de régulariser (Kuss & Graepel, 2003), soit de réduire la dimension

des matrices kernel (Hardoon *et al.*, 2004). Nous n'abordons pas dans ce travail la kernel CCA, mais le lecteur pourra se référer aux travaux de Vert & Kanehisa (2003) ou deYamanishi *et al.* (2003) pour l'application à des données de microarray.

CCA régularisée (rCCA). Le problème d'inversibilité de matrices peut être résolu en introduisant une contrainte de régularisation de type l_2 ou Ridge. Ce principe a été proposé dans le cadre de la CCA par Vinod (1976), puis Leurgans *et al.* (1993) pour des données fonctionnelles et enfin par González *et al.* (2008) dans le cadre de données de microarray. Le principe de la régularisation Ridge consiste à estimer les matrices de covariance $S_{XX} = \frac{1}{n}X'X$ et $S_{YY} = \frac{1}{n}Y'Y$ telles que

$$\begin{aligned} \hat{S}_{XX} = &\quad S_{XX} + \lambda_1 I_p \\ \hat{S}_{YY} = &\quad S_{YY} + \lambda_2 I_q \end{aligned}$$

avec $\lambda_1 \geq 0$ et $\lambda_2 \geq 0$, de telle façon que les matrices $\hat{S}_{XX}$ et $\hat{S}_{YY}$ soient régulières, et donc inversibles. On remplace ensuite S_{XX} et S_{YY} par $\hat{S}_{XX}$ et $\hat{S}_{YY}$ dans la CCA. Le réglage des paramètres (λ_1, λ_2) peut se faire par validation croisée afin d'obtenir des variables canoniques stables, comme proposé dans González *et al.* (2008). Cette approche semble donner des résultats biologiquement pertinents (Combes *et al.*, 2008) mais mériterait cependant une étude plus poussée concernant l'effet des paramètres de pénalisation sur les résultats. Par ailleurs, la version actuelle proposée dans le package `R cca` (González *et al.*, 2008) ne permet pas pour le moment de choisir les paramètres de régularisation pour chaque dimension CCA (ces paramètres sont réglés par une validation-croisée ou bien *leave-one-out* uniquement sur la première dimension CCA).

Greedy sparse CCA. Wiesel *et al.* (2008) proposent une sparse CCA basée sur une approche gourmande (*greedy*) avec une stratégie *forward*. Le problème à résoudre est le suivant :

$$\max_{u\neq 0, v\neq 0} \frac{u'S_{XY}v}{\sqrt{u'S_{XX}u}\sqrt{v'S_{YY}v}} \quad \text{sous les contraintes} \quad ||u||_0 \leq k_u \text{et} ||v||_0 \leq k_v \qquad (7.5)$$

avec $k_u < p$ et $k_v < q$ fixés. Cependant (7.5) est un problème combinatoire NP-difficile et une approche sous optimale est proposée, qui consiste tout simplement à choisir de façon séquentielle (sélection forward) les variables x^j et y^k telles que :

$$\max_{j,k} \frac{S_{XY}^{j,k}}{\sqrt{S_{XX}^{jj}}\sqrt{S_{YY}^{kk}}}$$

en utilisant des approximations de type "gourmand" que nous ne présenterons pas ici. Les auteurs montrent sur des données simulées de dimension raisonnable ($p + q = 2000$ et $n = 200$) que 80% de la corrélation CCA peut être prise en compte avec un quart des variables initiales. Cependant des applications sur des jeux de données réels manquent pour vraiment évaluer la pertinence de cette méthode intégrative. De plus,

on ne s'intéresse ici qu'à la première dimension canonique.
Notons qu'en posant $S_{XX}^{jj} = 1$ et $S_{YY}^{kk} = 1$ dans le problème (7.5), on peut de la même manière définir une sparse PLS (cf. section 7.3.2).

CCA Elastic Net (CCA-EN). Une approche qui introduit la pénalisation *Elastic Net* dans la CCA a été très récemment proposée par Waaijenborg *et al.* (2008). Les auteurs n'utilisent cependant pas la "vraie" formulation de la CCA, mais plutôt une version PLS avec mode de déflation canonique (présentée section 7.3.2), afin de se placer dans un contexte de régression. Elastic Net nécessite de régler ici 4 paramètres de pénalisation (λ_1 et λ_2 pour les normes l_1 et l_2 de chaque tableau), ce qui rend la tâche difficile dans les cas de grande dimension. À la place, et comme proposé par Zou & Hastie (2005) lorsque $p >> n$ dans la sparse PCA, les auteurs proposent de poser $\lambda_2 \to \infty$ pour un *univariate soft-thresholding*, ce qui conduit finalement à faire une pénalisation comme proposé dans Lê Cao *et al.* (2008), section 8. La méthode et l'agorithme sont détaillés dans la section 9.
Dans cette approche, les auteurs proposent de paramétrer les pénalisations lasso par validation croisée pour maximiser la corrélation entre les variables canoniques, comme dans González *et al.* (2008), mais pour chaque dimension CCA. Il en résulte des corrélations canoniques non décroissantes, puisque l'algorithme n'optimise pas le critère de corrélation ordinaire de la CCA, mais plutôt une version pénalisée de ce critère. En effet, on veut maximiser ici

$$\frac{v'X'Yu}{\sqrt{v'(I_p + \lambda_{1X}W_X^-)v}\sqrt{u'(I_q + \lambda_{1Y}W_Y^-)u}}, \tag{7.6}$$

où $W = diag|\hat{\beta}|$ et W^- est l'inverse généralisée de W, $\hat{\beta}$ étant l'estimateur lasso. Si $\lambda_{1X} = \lambda_{1Y} = 0$, alors on retombe sur un problème PLS de maximisation de covariance (cf. section 7.3.2). Cependant les auteurs initialisent leur algorithme (assez sensible à l'initialisation de départ) de façon à ce que les variables canoniques soient corrélées au maximum et que l'on converge vers une solution lasso unique.
Dans CCA-EN, il n'est pas possible de définir le nombre optimal de variables à sélectionner, et les auteurs proposent plutôt de choisir le nombre de variables à sélectionner dans chaque tableau. La méthode est évaluée sur des données simulées, puis sur les données *Glioma* de très grande dimension ($n = 45, p = 12210, q = 16872$), mais l'interprétation biologique des résultats n'est pas donnée.

Remarques. Il existe d'autres méthodes régularisées proches de la CCA pour les cas de données de grande dimension.
Nous pouvons citer l'Analyse de Co-Inertie (CIA, Doledec & Chessel, 1994; Dray *et al.*, 2003), qui s'apparente fortement à la CCA puisque les deux jeux de données sont pris en compte de façon symétrique. Le but de cette approche est d'identifier des structures similaires dans X et Y, de façon à maximiser la covariance. Le mode de résolution s'apparente fortement à la PLS-SVD (cf. section 7.3.2), bien que la décomposition des deux tableaux nécessite aussi le calcul de matrices diagonales résumant les correspondances

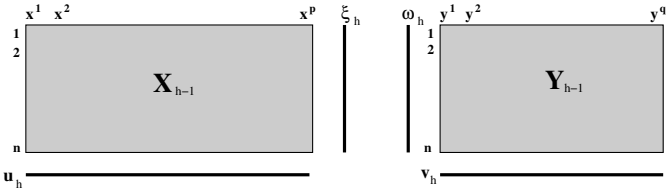


Fig. 7.2: PLS : décomposition en variables latentes (ξ_h, ω_h) et vecteurs loadings (u_h, v_h) des tableaux X et Y pour chaque dimension h de la PLS.

entre variables et échantillons. Comme dans la PLS, l'importance des variables dans le modèle est directement interprétable au niveau des composantes. Une application de la CIA a été présentée dans Culhane *et al.* (2003) dans le cas de comparaison d'expériences sur différentes plateformes génomiques (*cross-platform*) sur les mêmes individus pour les données NCI60 où des gènes différents ont été mesurés soit sur un support cDNA, soit sur un support Affymetrix. L'approche semble bien regrouper les individus étant affectés du même type de cancer et mettre en valeur des gènes complémentaires présents sur chaque support. Cependant, cette approche n'aura de sens que si les jeux de données X et Y sont de même nature.
Enfin notons que González (2008) a proposé une approche CCA sous contrainte lasso de la même manière que les approches SCoTLASS et DaLASS (Trendafilov & Jolliffe, 2006, 2007), cependant les premiers résultats s'inscrivent pour le moment sous la contrainte $n > p + q$ et se sont avérés très instables. Par ailleurs, nous avons aussi essayé des approches de sélection de variables en CCA sur la base de l'algorithme stochastique de la partie 2. Là encore, les résultats sont très instables. Ceci semble inhérent à la CCA qui considère les espaces engendrés par des variables pouvant être très corrélées entre elles ou encore dans des cas de problèmes très mal conditionnés.

7.3.2 Partial Least Squares regression

La régression Partial Least Squares (PLS) a été introduite par Wold (1966), tout d'abord sous forme de l'algorithme NIPALS, puis sous forme de nombreuses variantes. La PLS repose sur la décomposition simultanée des tableaux X et Y en vecteurs *loadings* et *variables latentes* associées (figure 7.2). L'idée principale étant que des régressions successives par projection sur des structures latentes permettent de mettre en évidence certains effets biologiques "cachés" (ou latents). Comme dans la PCA, les composantes PLS (variables latentes) sont des combinaisons linéaires des variables initiales. Cependant les coefficients définissant ces composantes ne sont pas linéaires puisqu'ils sont résolus via une succession de régressions locales sur ces composantes. Par ailleurs, la PLS va plus loin qu'un simple problème de régression, puisque l'on modèlise à la fois la structure de X *et* celle de Y simultanément par décompositions successives.
La fonction objectif à optimiser repose sur une maximisation de la covariance entre chaque combinaison linéaire des deux tableaux :

$$\arg \max_{u_h' u_h = 1, v_h' v_h = 1} cov(X u_h, Y v_h) \qquad h = 1 \dots H.$$

Les vecteurs loadings sont les vecteurs u_h et v_h pour chaque dimension PLS h et l'on notera $\xi_h = Xu_h$ et $\omega_h = Yv_h$ les variables latentes associées. Comme en PCA, les vecteurs loadings u_h et v_h peuvent être directement interprétés, puisqu'ils indiquent comment les variables x^j et y^k peuvent expliquer les relations entre X et Y. Les variables latentes ξ_h et ω_h contiennent plutôt des informations concernant les similarités ou dissimilarités entre individus ou échantillons (Wold *et al.*, 2004).
De très nombreuses variantes PLS existent suivant la forme des données et le but de la modélisation recherché : SIMPLS, "Straightforward Implementation of a statistically inspired Modification of the PLS method" (de Jong, 1993), PLS1 pour l'analyse univariée, PLS2 pour l'analyse multivariée, PLS-SVD pour résoudre le problème grâce à une décomposition SVD, kernel PLS pour résoudre de façon plus efficace le problème PLS lorsque $n << p + q$, etc. Afin de mieux comprendre l' article présenté section 8, nous présentons quelques-uns des algorithmes PLS de façon détaillée.

NIPALS. Cet algorithme a été le premier algorithme PLS, présenté par Wold (1966). Il permet notamment de faire une PCA du tableau X dans le cadre de données manquantes, sans avoir recours à la suppression ou l'estimation des valeurs manquantes. Nous ne présentons ici que l'algorithme sans données manquantes. Le lecteur pourra se référer à Tenenhaus (1998) pour la description de l'autre algorithme. Rappelons tout d'abord que pour des données centrées, la formule de décomposition de la PCA s'écrit

$$X = \sum_{h=1}^{H} \xi_h u_h'$$

où les vecteurs ξ_h sont les composantes principales et les vecteurs u_h les loadings ou vecteurs principaux, $h = 1 \dots H$. L'algorithme consiste en la construction d'une suite de tableaux $n \times p$ notés X_h de la façon suivante :

1. $X_0 = X$
2. Pour $h = 1 \dots H$:
   (a) ξ_h = première colonne de X_{h-1}
   (b) Répéter jusqu'à convergence de u_h :
   - $u_h = X_{h-1}'\xi_h / \xi_h'\xi_h$, normer u_h
   - $\xi_h = X_{h-1}u_h / u_h'u_h$
   (c) $X_h = X_{h-1} - \xi_h u_h'$

Nous adopterons par la suite la notation $\xi_h = X_{h-1}[, 1]$ plutôt que "ξ_h = première colonne de X_{h-1}". Notons que cette initialisation est arbitraire et indépendante du résultat final.
Avant normalisation du vecteur u_h, chaque élément de u_h représente le coefficient de régression de ξ_h lors de la régression des variables de x_{h-1}^j (les variables j de la matrice X_{h-1}) sur la composante ξ_h. De même, chaque élément de ξ_h représente le coefficient de régression de u_h lors de la régression des variables x_{h-1}^j sur u_h.

A la convergence de l'étape (b), les vecteurs ξ_h et u_h vérifient :

$$\begin{aligned} X'_{h-1}X_{h-1}u_h &= \lambda_h u_h \\ X_{h-1}X'_{h-1}\xi_h &= \lambda_h \xi_h \end{aligned}$$

où λ_h est la plus grande valeur propre de $X'_{h-1}X_{h-1}$ et $X_{h-1}X'_{h-1}$. On applique donc dans cet algorithme la méthode de la *puissance itérée* pour le calcul du vecteur propre d'une matrice associé à la plus grande valeur propre (Hotelling, 1936; Hoskuldsson, 1988).
L'étape (c), appelée aussi étape de *déflation*, calcule la matrice résiduelle X_h de la régression de X_{h-1} sur ξ_h. Ainsi, le problème de la PCA est résolu par une suite de régressions locales lors de l'étape (b), suivie d'une régression simple en (c).

PLS2. Passons maintenant à la régression PLS multivariée mettant en relation deux tableaux X et Y (Hoskuldsson, 1988). Les étapes de l'algorithme sont les suivantes :

1. $X_0 = X$, $Y_0 = Y$
2. Pour $h = 1 \dots H$:
   (a) $\xi_h = X_{h-1}[,1]$ $\quad$ $\omega_h = Y_{h-1}[,1]$
   (b) Répéter jusqu'à convergence de u_h :
      i. $u_h = X'_{h-1}\xi_h / \xi'_h\xi_h$, normer u_h
      ii. $\xi_h = X_{h-1}u_h / u'_h u_h$
      iii. $v_h = Y'_{h-1}\xi_h / \xi'_h\xi_h$, normer v_h
      iv. $\omega_h = Y_{h-1}v_h / v'_h v_h$
   (c) $c_h = X'_{h-1}\xi_h / \xi'_h\xi_h$ $\quad$ $d_h = Y'_{h-1}\xi_h / \xi'_h\xi_h$
   (d) $X_h = X_{h-1} - \xi_h c'_h$ $\quad$ $Y_h = Y_{h-1} - \xi_h d'_h$

Les sous-étapes (ii) et (iv) de l'étape (b) peuvent être considérées comme des régressions locales puisque l'on fait la régression de X_{h-1} sur ξ_h et Y_{h-1} sur ω_h pour obtenir des approximations de rang 1 des matrices (Abdi, 2003) :

$$\hat{X}_{h-1} = \xi_h(\xi'_h\xi_h)^{-1}\xi_h X_{h-1} \qquad \hat{Y}_{h-1} = \omega_h(\omega'_h\omega_h)^{-1}\omega_h Y_{h-1}.$$

Les coefficients de regression sont u_h et v_h :

$$u_h = \frac{X'_h\xi_h}{\xi'_h\xi_h} \qquad v_h = \frac{Y'_h\omega_h}{\omega'_h\omega_h}$$

et l'étape de déflation (d) peut aussi s'écrire :

$$X_h = X_{h-1} - \hat{X}_{h-1} \qquad Y_h = Y_{h-1} - \hat{Y}_{h-1}.$$

Ces dernières remarques permettent de comprendre que, par le biais de régressions locales successives, l'algorithme PLS2 ne nécessite pas le calcul de la matrice inverse de covariance et évite ainsi le problème des matrices mal conditionnées ou même non

inversibles rencontrées par la CCA.
Comme pour NIPALS, à la convergence de l'étape (b), les vecteurs u_h, v_h, ξ_h and ω_h vérifient les équations :

$$\begin{aligned} X'_{h-1}X_{h-1}Y'_{h-1}Y_{h-1}v_h &= \lambda_h v_h \\ X_{h-1}Y'_{h-1}Y_{h-1}X'_{h-1}\xi_h &= \lambda_h \xi_h \\ Y'_{h-1}Y_{h-1}X'_{h-1}X_{h-1}u_h &= \lambda_h u_h \\ Y_{h-1}X'_{h-1}X_{h-1}Y'_{h-1}\omega_h &= \lambda_h \omega_h \end{aligned} \tag{7.7}$$

L'étape cyclique (b) résout en fait le problème de recherche de vecteurs propres et valeurs propres via la méthode de la puissance itérée. Notons que la résolution d'une seule des équations ci-dessus suffirait à résoudre ce système en calculant ensuite les autres vecteurs propres via l'étape (b). D'après Hoskuldsson (1988), l'algorithme PLS2 converge en moins de 10 itérations.

PLS-SVD. La PLS-SVD (Lorber *et al.*, 1987) permet de résoudre le problème PLS de façon très efficace en décomposant la matrice $X'Y$ en valeurs et vecteurs singuliers. Notons que la décomposition SVD est une généralisation du problème de décomposition en valeurs propres et vecteurs propres (Golub & Van Loan, 1996). Cette technique permet notamment de résoudre d'autres problèmes d'analyse multivariée tels que la PCA, mais aussi en utilisant la SVD généralisée pour la CCA, ou l'Analyse Linéraire Discriminante (Jolliffe, 2002).
Pour la PLS, si on écrit $X'Y = U\Lambda V'$ avec U matrice de taille $p \times r$ et V matrice de taille $q \times r$, les vecteurs singuliers de gauche et droite (u_h et v_h, $h = 1 \dots H$, $H \leq r$) correspondent aux vecteurs loadings (Wegelin, 2000) et les valeurs singulières de la matrice Λ aux valeurs propres λ_h des relations (7.7). La déflation dans ce cas utilise l'approximation de rang 1 de la matrice $X'Y$, c'est-à-dire

$$(X'Y)_h = (X'Y)_{h-1} - \sum_{k=1}^{h-1} \lambda_k u_k v'_k.$$

L'avantage de cet algorithme est qu'il évite le calcul de la puissance itérée et qu'il est relativement efficace, la décomposition SVD n'étant calculée qu'une seule fois (Lorber *et al.*, 1987; Mevik & Wehrens, 2007). Le désavantage est que l'on risque de perdre l'orthogonalité entre les variables latentes de même type ξ_h et ω_h.
Notons par ailleurs que toute forme de variante de la PLS ne donne des résultats identiques que pour la première dimension.
La PLS-SVD a été utilisée par Nguyen & Rocke (2004) pour créer une PLS hybride. Elle a aussi été appliquée dans le cadre de données épidémiologiques (Sampson *et al.*, 1989; Streissguth *et al.*, 1993) sans que toutefois l'algorithme ait été explicité.
Dans la suite de nos travaux, nous nous sommes inspirés de cette variante PLS en raison de ses propriétés intéressantes : temps de calcul réduit et estimation des moindres carrés d'une matrice de rang inférieur par SVD.

Autres variantes. L'étape de déflation (d) de l'algorithme PLS2 sur Y_{h-1} est calculée selon le mode appelé *régression*. Ainsi, les deux matrices X_{h-1} et Y_{h-1} sont déflatées de manière *asymétrique* par rapport à ξ_h. On explicite ainsi la relation entre X et Y de façon asymétrique, puisque l'on cherche à expliquer Y en fonction de X.
Il existe aussi un mode de déflation appelé *canonique* (Tenenhaus (1998), ou "PLS-mode A", Wegelin, 2000) où les deux matrices X_{h-1} et Y_{h-1} sont déflatées de la manière suivante :

étape (c) $\quad c_h = X'_{h-1}\xi_h/\xi'_h\xi_h \qquad e_h = Y'_{h-1}\omega_h/\omega'_h\omega_h$
étape (d) $\quad X_h = X_{h-1} - \xi_h c'_h \qquad Y_h = Y_{h-1} - \omega_h e'_h$

On modélise ainsi une relation *symétrique* entre X et Y, de la même façon que pour CCA. En travaillant sur des jeux de données standardisés (centrés et réduits), Tenenhaus (1998) a montré que la PLS-mode A et la CCA donnent des résultats différents dans le cas $n < p + q$, bien que très proches. Aucune étude comparative n'a été faite encore concernant PLS-mode A et CIA, cependant nous donnons quelques éléments de réponse dans la section 9.

Modèles. Suivant le mode de déflation, on peut modéliser de deux façons différentes les espaces engendrés par X et Y. Par exemple, pour le mode régression (de Jong, 1993) :

$$X = \Xi C^T + \varepsilon_1 \qquad Y = \Xi D^T + \varepsilon_2 = XB + \varepsilon_2,$$

où Ξ est la matrice de taille $(n \times H)$ des variables latentes colonnes ξ_h et B $(p \times H)$ est la matrice des coefficients de regression. Les vecteurs colonnes des matrices C et D sont tels que $c_h = X^T_{h-1}\xi_h/(\xi'_h\xi_h)$ et $d_h = Y^T_{h-1}\xi_h/(\xi'_h\xi_h)$, et ε_1 $(n \times p)$ et ε_2 $(n \times q)$ sont les matrices résiduelles, $h = 1 \ldots H$.
Pour le mode canonique, les modèles deviennent :

$$X = \Xi C^T + \varepsilon_1 \qquad Y = \Omega E^T + \varepsilon_2,$$

où Ω est la matrice de taille $(n \times H)$ des variables latentes colonnes ω_h et les vecteurs colonnes de E sont définis tel que $e_h = Y^T_{h-1}\omega_h/(\omega'_h\omega_h)$, $h = 1 \ldots H$.

La PLS mode régression (PLS2) semble pour le moment la plus appliquée pour l'analyse de données biologiques (Boulesteix & Strimmer, 2005; Bylesjö *et al.*, 2007) en raison de l'objectif recherché. En effet le biologiste sait déjà quel type de variables influe sur l'autre groupe de variables. En général les variables y^k à prédire sont bien moins nombreuses que les variables x^j explicatives.
La PLS mode canonique, peu connue jusqu'à présent, permettrait en revanche de s'appliquer lorsque l'on n'a pas d'*a priori* sur l'explication de tel tableau par rapport à l'autre et que l'on aimerait adopter une approche exploratoire, ou s'il existe des relations réciproques entre les deux types de variables. Par exemple on pourrait comparer des mesures faite sur différents types de supports, comme des puces génomiques ADN

et affymetrix, pour des mêmes échantillons biologiques afin de regrouper des informations similaires ou complémentaires (voir section 9).
A part pour la dimension $h = 1$, où toutes les variantes PLS donnent des résultats identiques, ces deux modes de déflation donnent des résultats très différents dès que $h > 1$, puisque les matrices X_h et Y_h sont calculées différemment. Ceci explique les différences que nous avons observées dans nos travaux lorsque nous avons appliqué ces deux types de PLS.

7.4 Evaluation des méthodes

Comme nous l'avons évoqué dans la partie 2, il est difficile d'évaluer la pertinence de chacune des méthodes citées sans avoir recours à l'expertise biologique. En effet dans certains cas, la PLS2 par exemple, le mode régression permet d'évaluer la performance prédictive du modèle grâce à la validation croisée. C'est généralement l'approche qui a été conseillée (Tenenhaus, 1998; Boulesteix & Strimmer, 2005; Mevik & Wehrens, 2007).
En revanche pour les méthodes canoniques (CCA, PLS-mode A) l'évaluation de la performance de l'approche est un problème qui a été peu étudié. Pour une approche type sparse CCA, on peut évaluer la corrélation des variables canoniques en fonction de la régularisation appliquée (González, 2008; Wiesel *et al.*, 2008). Mais pour une CCA classique ou une PLS-mode A, il reste à comparer ces méthodes entres elles grâce à certains critères. Tenenhaus (1998) propose par exemple de calculer la part de variance expliquée d'un tableau (*e.g.* X) par la composante du même groupe (*e.g.* ξ_h avec les notations PLS) de la façon suivante en calculant le critère de "redondance" Rd :

$$Rd(X, \xi_h) = \frac{1}{p} \sum_{j=1}^{p} cor^2(x^j, \xi_h).$$

De la même manière on peut calculer la part de variance expliquée d'un tableau avec les composantes de l'autre groupe (*e.g* ω_h).
Enfin pour les approches sparse PCA, les auteurs cités ont essentiellement comparé la variance expliquée par chaque composante principale par rapport à une PCA classique. Tous ces critères statistiques restent cependant imprécis ou insatisfaisant compte tenu des caractéristiques de nos données. Par conséquent il est très important de considérer la pertinence biologique des résultats pour convaincre le biologiste.

De manière générale, il convient maintenant de confronter ces approches régularisées sur de vrais jeux de données adaptés aux approches, en donnant une interprétation détaillée des résultats dans le domaine d'expertise correspondant. Attention, l'application de telle ou telle méthode doit dépendre de l'objectif biologique fixé au départ. Certains articles commencent à prendre en compte cet aspect des choses. C'est le cas de Bylesjö *et al.* (2007) qui appliquent une variante PLS (appelée 02-PLS, Trygg & Wold, 2003) chez les plantes *Populus* pour intégrer des données de métabolites et de transcrits. Ils comparent ensuite leurs résultats avec une PCA des deux jeux de

données concaténés et standardisés. Nous pouvons aussi citer les travaux de González *et al.* (2008); Combes *et al.* (2008); González (2008) dans le cas de la ridge CCA, où un accent particulier est donné sur l'interprétation biologique des résultats pour présenter cette approche.

7.5 Plan de la partie

Dans cette partie, nous présentons deux articles basés sur le développement et l'application d'une méthode d'intégration nommée *sparse PLS*, qui permet la sélection de variables omiques issues de deux tableaux. Dans le premier travail, nous avons voulu montrer la pertinence de l'approche, par rapport à une PLS classique, à la fois sur des critères statistiques et biologiques. Dans le deuxième travail, nous nous sommes surtout focalisés sur le mode canonique proposé dans la sparse PLS, en comparant sur un même jeu de données les résultats biologiques obtenus avec d'autres méthodes canoniques parcimonieuses ou régularisées : analyse de Co-Inertie et CCA-Elastic Net. Nous proposons ensuite quelques éléments de discussion et perspectives concernant cette deuxième partie.

8. Article méthodologique

Cet article présente l'approche sparse PLS, dérivée d'une variante PLS-SVD, dans laquelle une pénalisation l_1 est appliquée. Cette approche permet de sélectionner des variables de types différents dans un contexte d'intégration de données. Nous montrons que l'approche conserve les bonnes propriétés de la PLS sur plusieurs jeux de données publiques. Sur un des jeux de données, nous montrons le gain de la sparse PLS par rapport à la PLS, grâce à une interprétation biologique détaillée.
Dans cette approche, nous proposons deux schémas de déflation (mode régression ou canonique) suivant la nature du problème biologique. Nous proposons quelques critères pour paramétrer les deux pénalisations et choisir le nombre de composantes PLS. Ces critères (validation croisée par exemple) sont cependant vite limités en raison du petit nombre d'échantillons étudiés, et nécessitent plutôt l'avis du biologiste en terme de nombre de variables à sélectionner.
Encore une fois la validation de l'approche est difficile compte tenu du petit nombre d'échantillons. Ce travail montre l'importance véritable d'une collaboration entre statisticiens et biologistes afin de montrer la pertinence des résultats.

Cet article est en seconde lecture dans la revue Statistical Application in Genetics and Molecular Biology. L'application biologique de l'approche développée résulte d'une collaboration avec D. Roussouw lors de mon séjour à l'Université de Cape Town.

A Sparse PLS for Variable Selection when Integrating Omics data

Kim-Anh Lê Cao[1,2], Debra Rossouw[3], Christèle Robert-Granié[2] and Philippe Besse[1]

Abstract

Recent biotechnology advances allow for the collection of multiple types of omics data sets, such as transcriptomic, proteomic or metabolomic data to be integrated. The problem of feature selection has been addressed several times in the context of classification, but has to be handled in a specific manner when integrating data. In this study, we focus on the integration of two-block data sets that are measured on the same samples. Our goal is to combine integration and simultaneous variable selection on the two data sets in a one-step procedure using a PLS variant to facilitate the biologists interpretation. A novel computational methodology called "sparse PLS" is introduced for a predictive purpose analysis to deal with these newly arisen problems. The sparsity of our approach is obtained by soft-thresholding penalization of the loading vectors during the SVD decomposition.
Sparse PLS is shown to be effective and biologically meaningful. Comparisons with classical PLS are performed on simulated and real data sets and a thorough biological interpretation of the results obtained on one data set is provided. We show that sparse PLS provides a valuable variable selection tool for high dimensional data sets.

Introduction

Motivation. Recent advances in technology enable the monitoring of an unlimited quantity of data from various sources. These data are gathered from different analytical platforms and allow their integration among different types, such as transcriptomic, proteomic or metabolomic data. This integrative biology approach enables to understand better some underlying biological mechanisms and interaction between functional levels, if one succeeds in incorporating the several omics types of data, that are characterized by many variables but not necessarily many samples or observations. In this

[1]Institut de Mathématiques, Université de Toulouse et CNRS (UMR 5219), F-31062 Toulouse, France
[2]Station d'Amélioration Génétique des Animaux UR 631, Institut National de la Recherche Agronomique, F-31326 Castanet, France
[3]Institute for Wine Biotechnology, University of Stellenbosch, Stellenbosch, South Africa

highly dimensional setting, the selection of genes, proteins or metabolites is absolutely crucial to overcome computational limits (from a mathematical and statistical point of view) and to facilitate the biological interpretation. Hence our quest of sparsity is motivated by the biologists needs, who want to separate the useful information related to the study from the non useful information, due to experiment inaccuracies. The resulting variable selection might also enable a feasible biological validation with a reduced experimental cost. We especially focus on the integration context, which is the main goal of omics data. For example, one biological study might aim at explaining the q metabolites by the p transcripts, that are measured on the same n samples. In this typical case, $n << p + q$.

In this study, we propose a sparse version of the PLS, that aims at combining selection *and* modelling in a one-step procedure for such problems. Our sparse PLS is based on soft-thresholding penalization and is obtained by penalizing a sparse SVD (Shen and Huang, 2007), using a hybrid PLS with SVD decomposition (Lorber et al., 1987). This approach deals with integration problems, that cannot be solved with usual feature selection approaches proposed in classification or discrimination studies where there is only one data set to analyse. Hence, multiple testing that looks for differentially expressed genes does not apply here, as well as other classification methods that were applied to transcriptomic data sets. In this latter case, many authors (among them: Guyon et al. 2002; Lê Cao et al. 2007) have applied feature selection methods to microarray data and have been proved to bring biologically meaningful genes lists. However, in our context, the feature selection aim has to be integrated with modelling, and very few approaches have been proposed to deal with these newly arisen problems, especially in a one-step procedure. In a two-block data sets setup, our aim is to *predict* one group of variables from the other group. Several approaches that seek linear combinations of both groups of variables can answer this biological problem. However, they are often limited by collinearity or ill posed problems, that require regularization techniques, such as l_1 (soft-thresholding, Lasso) or l_2 (Ridge) penalizations.

Background and related work. Partial Least Squares regression (PLS, Wold 1966) is a well known regression technique, mostly applied in chemometrics. Its stability property faced to collinear matrices gives PLS a clear superiority to CCA, multiple linear regression, ridge regression or other regression techniques. Furthermore, since Wold original approach, many variants have arisen (SIMPLS, de Jong 1993, PLS1 and 2, PLS-A, PLS-SVD, see Wegelin (2000) for a survey) that provide the user a solution for almost any problem. We will describe and discuss some of these variants in this study.

PLS has been successfully applied to biological data, such as gene expression (Datta, 2001), integration of gene expression and clinical data (with bridge PLS, Gidskehaug

et al. 2007), integration of gene expression and ChIP connectivity data (Boulesteix and Strimmer, 2005) and more recently for reconstructing interaction networks from microarray data (Pihur et al., 2008). We can also mention the study of (Culhane et al., 2003) who applied Co-Inertia Analysis (CIA, Doledec and Chessel 1994) from which PLS is a particular case, in a cross platform comparison in microarray data.

In the context of feature selection from both data sets, one closely related work proved to bring biologically meaningful results is the O2PLS model (Trygg and Wold, 2003), associated to variable selection in Bylesjö et al. (2007) for combining and selecting transcript and metabolite data in *Arabidopsis Thaliana* in a regression framework. O2PLS decomposes each data set in three structures (predictive, unique and residual). The most dominating correlation and covariance in both sample directions and variable directions is extracted and can be interpreted. Variable selection is then performed on the correlation loadings with a permutation strategy, hence with a two-step procedure.

More recently, Waaijenborg et al. (2008) and Chun and Keles (2007) both adapted Elastic Net regularization (Zou and Hastie, 2005) in the PLS, either in a canonical framework, or in a regression framework, by directly penalizing the optimization problem. Both approaches seem promising, as Chun and Keles (2007) demonstrated that the PLS consistency property does not hold when $n << p + q$. However, it would be useful to show the biological relevancy of their results. Nevertheless, their studies show the need of developing such integrative methods for biological problems.

Our contribution and results. We propose a sparse PLS approach that combines both integration and variable selection simultaneously on the two data sets, in a one-step strategy. We show that our approach is applicable on high-throughput data sets and bring more relevant results compared to PLS.

Outline of the paper. A brief introduction to PLS will be given, before describing the sparse PLS method. We detail how to add sparsity to PLS with a soft-thresholding penalization combined to SVD computation (Shen and Huang, 2007). We then assess the validity of the approach on one simulated and three real data sets, compare and discuss the results with a classical PLS approach. We also provide a full biological interpretation of the results obtained on a typical integrative study of wine yeast, that combines transcripts and metabolites. We show how sparse PLS highlights the most essential transcripts that are meaningfully related to the metabolites.

1 Methods

1.1 PLS

The PLS regression looks for a decomposition of centered (possibly standardized) data matrices X ($n \times p$) and Y ($n \times q$) in terms of components scores, also called latent variables: $(\xi_1, \xi_2 \dots \xi_H)$, $(\omega_1, \omega_2 \dots \omega_H)$, that are n-dimensional vectors, and associated loadings: $(u_1, u_2 \dots u_H)$, $(v_1, v_2 \dots v_H)$, that are respectively p and q- dimensional vectors, to solve the following optimization problem (Burnham et al., 1996):

$$\max_{||u_h||=1, ||v_h||=1} cov(X_{h-1}u_h, Yv_h) \tag{1}$$

where X_{h-1} is the residual (deflated) X matrix for each PLS component dimension $h = 1 \dots H$. Problem (1) is equivalent to solve: $\max cov(\xi_h, \omega_h)$.

Many PLS variants exist depending on the way X and Y are deflated, either in a symmetric ("PLS-mode A") or asymmetric way ("PLS2") (Tenenhaus, 1998; Wegelin, 2000), and the models will consequently differ. In this study, we will focus only on a regression framework.

In the case of a *regression mode* (asymmetric), the models of X- and Y-space are respectively (Hoskuldsson, 1988):

$$X = \Xi C^T + \varepsilon_1 \qquad Y = \Xi D^T + \varepsilon_2 = XB + \varepsilon_2 \tag{2}$$

where Ξ $(n \times H)$ is the matrix of PLS components ξ_h, B $(p \times H)$ is the matrix of regression coefficients. The column vectors of C and D are defined as $c_h = X_{h-1}^T \xi_h / (\xi_h' \xi_h)$ and $d_h = Y_{h-1}^T \xi_h / (\xi_h' \xi_h)$, and ε_1 $(n \times p)$ and ε_2 $(n \times q)$ are the residual matrices, $h = 1 \dots H$.

Another PLS alternatives exist depending if X and Y are deflated separately or directly using the cross product $M = X^T Y$ and the SVD decomposition. We will discuss these various approaches in sections 1.2 and 1.4. Note that in any case, all PLS variants are equivalent during the computation of the first dimension.

1.2 SVD decomposition and PLS-SVD

We recall the SVD decomposition and the principle of the PLS-SVD approach, that will be useful for understanding our sparse PLS approach.

1.2.1 Singular Value Decomposition

Any real r-rank matrix M $(p \times q)$ can decomposed into three matrices U, Δ, V as follows:

$$M = U \Delta V^T$$

where $U(p \times r)$ and $V(q \times r)$ are orthonormal and $\Delta(r \times r)$ is a diagonal matrix whose diagonal elements δ_k $(k = 1 \dots r)$ are called the singular values. The singular values are equal to the square root of the eigenvalues of the matrices $M^T M$ and MM^T. One interesting property that will be used in our sparse PLS method is that the columns vectors of U and V, noted $(u_1, \dots, u_r)$ and $(v_1, \dots, v_r)$ (resp. called left and right singular vectors) correspond to the PLS loadings of X and Y if $M = X^T Y$.

1.2.2 PLS-SVD

In PLS-SVD, the SVD decomposition of $M = X^T Y$ is performed only once, and for each dimension h, M is directly deflated by its rank-one approximation ($M_h = M_{h-1} - \delta_h u_h v_h'$). This computationally attractive approach may however lead to non mutually orthogonal latent variables, a property of PLS2 ($\xi_s' \xi_r = 0, r < s$) and PLS-mode A ($\xi_s' \xi_r = 0$ *and* $\omega_s' \omega_r = 0, r < s$).

1.3 Soft-thresholding penalization

Shen and Huang (2007) proposed a sparse PCA approach using the SVD decomposition of $X = U \Delta V^T$ by penalizing the PCA loading vector v_k. The optimization problem to solve is

$$\min_{u,v} ||X - uv'||_F^2 + P_\lambda(v) \tag{3}$$

where $||X - uv'||_F^2 = \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - u_i v_j)^2$ and $P_\lambda(v) = \sum_{j=1}^p p_\lambda(|v_j|)$ is a penalty function. Among different penalty functions that were proposed, we considered the soft-thresholding function.

Solving (3) is performed in an iterative way, as described below:

- Decompose $X = U \Delta V^T$, $X_0 = X$
- For h in 1..H:
  1. Set $v_{old} = \delta_h v_h^\star$, $u_{old} = u_h^\star$, where $v_h^\star$ and $v_h^\star$ are unit vectors
  2. Until convergence of u_{new} and v_{new}:
     (a) $v_{new} = g_\lambda(X_{h-1}^T u_{old})$
     (b) $u_{new} = X^T v_{new} / ||X_{h-1}^T v_{new}||$
     (c) $u_{old} = u_{new}$, $v_{old} = v_{new}$
  3. $v_{new} = v_{new} / ||v_{new}||$
  4. $X_h = X_{h-1} - \delta_h u_{new} v_{new}'$

where $g(y) = sign(y)(|y| - \lambda)_+$ is the soft-thresholding penalty function.

In our particular PLS case, we are interested in penalizing both loadings vectors u_k

and v_k to perform variable selection in both data sets. Indeed, one interesting property of PLS is the direct interpretability of the loadings vectors as a measure of the relative importance of the variables in the model (Wold et al., 2004). Our optimization problem becomes:

$$\min_{u,v} ||M - uv'||_F^2 + g_{\lambda_1}(u) + g_{\lambda_2}(v) \tag{4}$$

which is solved iteratively by replacing X by M and the steps 2.**a.** and 2.**b.** by:

$$v_{new} = g_{\lambda_1}(M_{h-1}^T u_{old})$$

$$u_{new} = g_{\lambda_2}(M_{h-1} v_{old})$$

The sparse PLS algorithm is detailed in next section.

1.4 Sparse PLS

It is easy to understand that during the deflation step of the PLS-SVD, $M_h \neq X_h^T Y_h$. This is why we propose to compute separately X_h and Y_h, then to decompose at each step $\tilde{M}_h = X_h^T Y_h$ and finally, to extract the first pair of singular vectors. As Hoskuldsson (1988) explains, taking one pair of loadings (u_h, v_h) at a time will lead to a biggest reduction of the total variation in the X and Y-spaces. In our approach, the SVD decomposition will provide a useful tool for selecting variables from each of the two-blocks data. We now detail the sparse PLS algorithm (*sPLS*) based on the iterative PLS algorithm (see Tenenhaus 1998) and SVD computation of $\tilde{M}_h$ for each dimension.

1. $X_0 = X \qquad Y_0 = Y$
2. For h in 1..H:
   (a) Set $\tilde{M}_{h-1} = X_{h-1}^T Y_{h-1}$
   (b) Decompose $\tilde{M}_{h-1}$ and extract the first pair of singular vectors $u_{old} = u_h$ and $v_{old} = v_h$
   (c) Until convergence of u_{new} and v_{new}:
      i. $u_{new} = g_{\lambda_2}(\tilde{M}_{h-1} v_{old})$, norm u_{new}
      ii. $v_{new} = g_{\lambda_1}(\tilde{M}_{h-1}^T u_{old})$, norm v_{new}
      iii. $u_{old} = u_{new}$, $v_{old} = v_{new}$
   (d) $\xi_h = X_{h-1} u_{new} / u'_{new} u_{new}$
   $\omega_h = Y_{h-1} v_{new} / v'_{new} v_{new}$
   (e) $c_h = X_{h-1}^T \xi_h / \xi'_h \xi_h$
   $d_h = Y_{h-1}^T \xi_h / \xi'_h \omega_h$
   $e_h = Y_{h-1}^T \omega_h / \omega'_h \omega_h$

(f) $X_h = X_{h-1} - \xi_h c'_h$

(g) $Y_h = Y_{h-1} - \xi_h d'_h$

Note that in the case where there is no sparsity constraint ($\lambda_1 = \lambda_2 = 0$) we obtain the same results as in a classical PLS.

1.5 Missing data

When dealing with biological data, it is very common to be confronted to missing data. In order not too lose too much information, an interesting approach to substitute each missing data with a value can be the Non LInear Estimation by Iterative Partial Least Squares (NIPALS, Wold 1966). This method has been at the origin of PLS and allows performing PCA with missing data on each block data set. Details of the algorithm can be found in Tenenhaus (1998). Several studies show that the convergence of NIPALS and its good estimation are limited by the number of missing values (20-30%), see for example Dray et al. (2003). NIPALS is now implemented in the `ade4` package.

1.6 Tuning criteria and evaluation

1.6.1 Soft-thresholding penalization

The two penalization parameters λ_1^h and λ_2^h can be simultaneously chosen by computing the error prediction ("*RMSEP*" see section 1.6.3) with k-fold cross validation or leave-one-out cross validation, and this for each given dimension h. In practice however, when analyzing biological data, our experience showed that an optimal tuning of the penalization parameters by optimizing the predictive ability of the model, does not necessarily satisfy the biologists needs. Indeed, in biological data sets, many omics data are still unknown (*e.g* associated functions, annotations) and too small variable selections might not allow for the biologists to correctly assess the results. This is why they may prefer instead to choose the number of non zero components in each loading vector u_h, v_h or in both, for each dimension h. This option was proposed in Zou and Hastie (2005) in their `R` package `elasticnet` for their sparse PCA.

1.6.2 Choice of PLS dimension

Marginal contribution of the latent variable ξ_h. In the case of a regression context, Tenenhaus (1998) proposed to compute a criteria called Q_h^2 that measures the marginal contribution of ξ_h to the predictive power of the PLS model, by performing cross validation computations. Here, as the number of samples n is usually small, we propose to use leave-one-out cross validation (loo-cv). Q_h^2 is computed for all Y variables and is defined as

$$Q_h^2 = 1 - \frac{\sum_{k=1}^{q} PRESS_{kh}}{\sum_{k=1}^{q} RSS_{k(h-1)}},$$

where $PRESS_h^k = \sum_{i=1}^{n}(y_i^k - \hat{y}_{h(-i)}^k)^2$ is the PRediction Error Sum of Squares and $RSS_h^k = \sum_{i=1}^{n}(y_i^k - \hat{y}_{hi}^k)^2$ is the Residual Sum of Squares for the variable k and the PLS dimension h.

We define the estimated matrix of regression coefficients $\hat{B}$ of B, using the same notation as in equation (2): $\hat{B} = U^* D^T$ where $U^* = U(C^T U)^{-1}$ (see De Jong and Ter Braak 1994; Tenenhaus 1998) and where the column vectors of U are the loading vectors $(u_1, \ldots, u_h)$, $h = 1 \ldots H$. For any i sample, we can predict $\hat{y}_{hi}^k = x_{hi}\hat{B}_{h(-i)}^k$.

This criteria is the one adopted in the *SIMCA-P* software (developed by S. Wold and Umetri 1996). The rule to decide if ξ_h contributes significantly to the prediction is if

$$Q_h^2 \geq (1 - 0.95^2) = 0.0975$$

However, the choice of the PLS dimension still remains an open question that has been mentioned by several authors (see Mevik and Wehrens 2007; Boulesteix 2004). In our particular biological context, we can show that graphical representations of the samples facilitate this choice as the plots of (ξ_h, ξ_{h+1}) and (ω_h, ω_{h+1}) do not have a biological meaning if h is too large. In fact, our results (see below) show that all relevant information in terms of identification of the measured biological effects can be extracted from 3 dimensions.

1.6.3 Evaluation

RMSEP For a regression context, Mevik and Wehrens (2007); Boulesteix (2004) in the `R` `pls` and `plsgenomics` packages proposed to compute the Root Mean Squared Error Prediction criterion (RMSEP) with cross validation in order to choose the H parameter. As we already suggested to use the Q_h^2 criterion for this issue, we propose instead to use the RMSEP criterion as a way of evaluating the predictive power for each Y variable between the original non-penalized PLS and the sPLS in the next section.

Note that the Q_h^2 criteria is closely related to RMSEP ($= PRESS_{kh}$) and gives a more general insight of the PLS, whereas the RMSEP requires to be computed for each variable k in Y.

2 Validation studies

The evaluation of any statistical approach is usually performed with simulated data sets. In the context of biological data, however, simulation is a difficult exercise as one has to take into account technical effects that are not even easily identifiable on the real data sets. We first propose to simulate as realistically as possible two-block data

sets in a regression framework, to answer the questions : does the sparse PLS select relevant variables ? Does the variable selection performed simultaneously on both data sets improve the predictive ability of the model, compared to the PLS that includes all variables in the model ? Once these questions are answered, we propose on the next step to show that our approach is applicable on biological data sets with various complexities, and that it may give potentially relevant results from a statistical point of view compared to PLS. Finally, in the next section, we provide a detailed biological interpretation for one of the data set, and show that sparse PLS answers the biological question compared to the PLS.

2.1 Simulation study

2.1.1 Simulation design

As proposed by Chun and Keles (2007), this simulation is designed to compare the prediction performance of the PLS and sPLS in the case where the relevant variables are not governed by a latent variable model. In this setting, we also added two cross conditions to complexify this setting. We set $p = 5000$ genes, $q = 50$ response variables and $n = 40$ samples, all with base error model being Gaussian with unit variance. We defined the mean vectors μ_1 and μ_2 as follows and divided the samples into consecutive blocks of 10, denoted by the sets (a, b, c, d), where

$$\mu_{1i} = \begin{cases} -2 & \text{if } i \in a \cup b \\ +2 & \text{otherwise.} \end{cases}$$

$$\mu_{2i} = \begin{cases} -1.5 & \text{if } i \in a \cup c \\ +1.5 & \text{otherwise.} \end{cases}$$

For the first 20 genes, we generated 20 columns of X from a multivariate normal with an AR(1) covariance matrix with auto correlation $\rho = 0.9$. These genes will get a strong Y response, but should not be of interest in the model. The next 40 genes have the mean structure μ_1 or μ_2:

$$x_{ij} = \mu_{1i} + \epsilon_{ij}, \quad j = 21 \ldots 40, \quad i = 1 \ldots n.$$

$$x_{ij} = \mu_{2i} + \epsilon_{ij}, \quad j = 41 \ldots 60, \quad i = 1 \ldots n.$$

The next genes have the mean structure U_m and are generated by $X_j = U_m + \epsilon_j$, $m = 1 \ldots 4$,

$$\begin{aligned} U_1 = &\quad -1.5 + 1.5\mathbb{1}_{u_{ij} \leq 0.4}, \quad 1 \leq i \leq n, \quad 61 \leq j \leq 80, \\ U_2 = &\quad +1.5 - 1.5\mathbb{1}_{u_{ij} \leq 0.7}, \quad 1 \leq i \leq n, \quad 81 \leq j \leq 100, \\ U_3 = &\quad -2 + 2\mathbb{1}_{u_{ij} \leq 0.3}, \quad 1 \leq i \leq n, \quad 101 \leq j \leq 120, \\ U_4 = &\quad +2 - 2\mathbb{1}_{u_{ij} \leq 0.3}, \quad 1 \leq i \leq n, \quad 121 \leq j \leq 140, \end{aligned}$$

Table 1: Averaged RMSEP (standard error) for each PLS dimension for 50 simulated data sets.

	PLS	sparse PLS
dim 1	0.930 (0.009)	0.715 (0.030)
dim 2	0.927 (0.009)	0.581 (0.019)
dim 3	0.926 (0.009)	0.580 (0.019)

where $u_{ij} \sim \mathcal{U}(0,1)$ and ϵ_j are i.i.d random vectors from $\mathcal{N}(0, \mathbb{1}_n)$. In all cases, $\epsilon_{ij} \sim \mathcal{N}(0,1)$, which is also how the remaining 4860 genes are defined.

The response variables Y_{ik} follow $Y_k = X\beta_1 + e_k$, $k = 1 \dots 10$, with

$$\beta_{1j} = \begin{cases} 10 & \text{if } 1 \leq j \leq 20 \\ 8 & \text{if } 21 \leq j \leq 40 \\ 4 & \text{if } 21 \leq j \leq p, \end{cases}$$

and $Y_k = X\beta_2 + e_k$, $k = 11 \dots 20$ with

$$\beta_{2j} = \begin{cases} 10 & \text{if } 1 \leq j \leq 20 \\ 4 & \text{if } 21 \leq j \leq 40 \\ 8 & \text{if } 21 \leq j \leq p \end{cases}$$

and $Y_k \sim e_k$ for $k = 21 \dots 50$ with $e_k \sim \mathcal{N}(0, \mathbb{1}_n)$.

In this simulation setting, the tested methods should highlight the genes X_j, $j = 11 \dots 40$ and the response variables Y_k, $k = 1 \dots 30$, which are related either to a μ_1 or μ_2 effect.

2.1.2 Prediction performance

X and Y are simulated 50 times and we use 10-fold cross validation on each data set. For the sparse PLS, we arbitrarily chose to select 50 genes and 30 response variables for each dimension h, $h = 1 \dots 3$. For PLS, no penalization is applied, so that all Y variables are modelled with respect to the whole X data set for each simulation run.

The RMSEP for each response variable, each test set and each dimension is computed and averaged in Table 1. These first results show that sparse PLS improves the predictive ability of the model. After dimension $H = 2$, neither sPLS nor PLS get a significantly decreasing averaged RMSEP. This is in agreement with our simulation design, in which only two latent effects, the μ_1 and μ_2 effects, are included. The next section show that these effects are indeed highlighted by PLS and sPLS in the first 2 dimensions.

2.1.3 Variable selection

In this part, we compare the loading vectors (u_1, u_2, u_3) and (v_1, v_2, v_3) in the PLS and the sPLS in one simulation run (results were similar for the other runs). Figure 1 shows

Figure 1: Absolute variable weights in the loading vectors of PLS (left) or sparse PLS (right) for the first 100 X variables (top) and the Y variables (bottom). The whole X variables weights can be found in supplementary material. Red (green) color stands for the variables related to the μ_1 (μ_2) effect.

Table 2: Description of the data sets.

	Liver Toxicity	Arabidopsis	Wine Yeast
# samples n	64	18	43
X	gene expression	transcript	transcript
p	3116	22 810	3381
Missing values	2	0	0
Y	clinic variables	metabolite	metabolite
q	10	137	22
Missing values	0	22	0

that both PLS and sparse PLS highlight the "good" genes, but with no clear distinction between the group of genes with a μ_1 or μ_2 effect for the PLS in dimension 1 or 2. On the contrary, the sparse PLS clearly selects the μ_1 effect genes on dimension 2 with heavy weights. This may be useful for the biologists who want to clearly separate the genes related to each effect on a different dimension. For both methods, the dimension 3 does not seem to be informative. The same conclusion can be drawn for the Y variables.

If an artificial two step selection procedure is performed in PLS, first by ordering the absolute values of the loadings and then selecting a chosen number of variables, to select 50(30) genes (response variables) for the first three dimensions, the two selections in PLS and sPLS are roughly the same (identical for dimension 1, up to 5 different selected variables in dimension 2 and 3). This shows that sPLS simply seems to shrink the PLS loading coefficients in this simple controlled setting. However, on real data sets (see below), the difference between the two methods is genuine in terms of variable selection.

2.2 Case studies

2.2.1 Data sets

Liver Toxicity study In the liver toxicity study (Heinloth et al., 2004), 4 male rats of the inbred strain Fisher 344 were exposed to non-toxic (50 or 150 mg/kg), moderately toxic (1500 mg/kg) or severely toxic (2000 mg/kg) dose of acetaminophen (paracetamol) in a controlled experiment. Necropsies were performed at 6, 18, 24 and 48 hours after exposure and the mRNA from the liver was extracted. Ten clinical chemistry measurements variables containing markers for liver injury are available for each object and numerically measure the serum enzymes level. The expression data are arranged in a matrix X of $n = 64$ objects and $p = 3116$ expression levels after normalization and pre-processing (Bushel et al., 2007). There are 2 missing values in the gene expression matrix.

In the original descriptive study, the authors claim that the clinical variables might

not help detecting the paracetamol toxicity in the liver, but that the gene expression could be an alternative solution. However, in a PLS framework, we can be tempted to predict the clinical parameters (Y) by the gene expression matrix (X), as performed in Gidskehaug et al. (2007).

Arabidopsis data The responses of 22810 transcript levels and 137 metabolites and enzymes (including 67 unidentified metabolites) during the diurnal cycle (6) and an extended dark treatment (6) in WT Arabidopsis, and during the diurnal cycle (6) in starch less pgm mutants, is studied (Gibon et al., 2006). The aim is to detect the change of enzyme activities by integrating the changes in transcript levels and detect the correlation between the different time points and the 3 genotypes.

According to this previous study, metabolites and enzymes are regulated by gene expressions rather than vice versa. We hence assigned to the X matrix the transcript levels and to the Y matrix the metabolites. The Y data set contained 22 missing values. This data set is characterized by a very small number of samples (18).

Wine Yeast data set *Saccharomyces cerevisiae* is an important component of the wine fermentation process and determines various attributes of the final product. One such attribute that is important from an industrial wine-making perspective is the production of volatile aroma compounds such as higher alcohols and their corresponding esters (Nykanen and Nykanen, 1977; Dickinson et al., 2003). The pathways for the production of these compounds are not clearly delineated and much remains unknown regarding the roles and kinetics of specific enzymes. In addition, most of the key reactions in the various pathways are reversible and the enzymes involved are fairly promiscuous regarding substrate specificity (Bely et al., 1990; Ribéreau-Gayon et al., 2000). In fact, different yeast strains produce wines with highly divergent aroma profiles. The underlying genetic and regulatory mechanisms responsible for these differences are largely unknown due to the complex network structure of aroma-producing reactions. As such an unbiased, holistic systems biology approach is a powerful tool to mine and interpret gene expression data in the context of aroma compound production.

In this study, five different industrial wine yeast strains (VIN13, EC1118, BM45, 285, DV10) were used in fermentation with synthetic must, in duplicate or triplicate (biological repeats). Samples were taken for microarray analysis at three key time points during fermentation, namely Day2 (exponential growth phase), Day5 (early stationary phase) and Day14 (later stationary phase). Exometabolites (aroma compounds) were also analysed at the same time by GC-FID.

Microarray analysis was carried out using the Affymetrix platform, and all normalizations and processing was performed according to standard Affymetrix procedures. To compensate for the bias towards cell-cycle related genes in the transcriptomic data set, the data was pre-processed to remove genes that are exclusively involved in cell

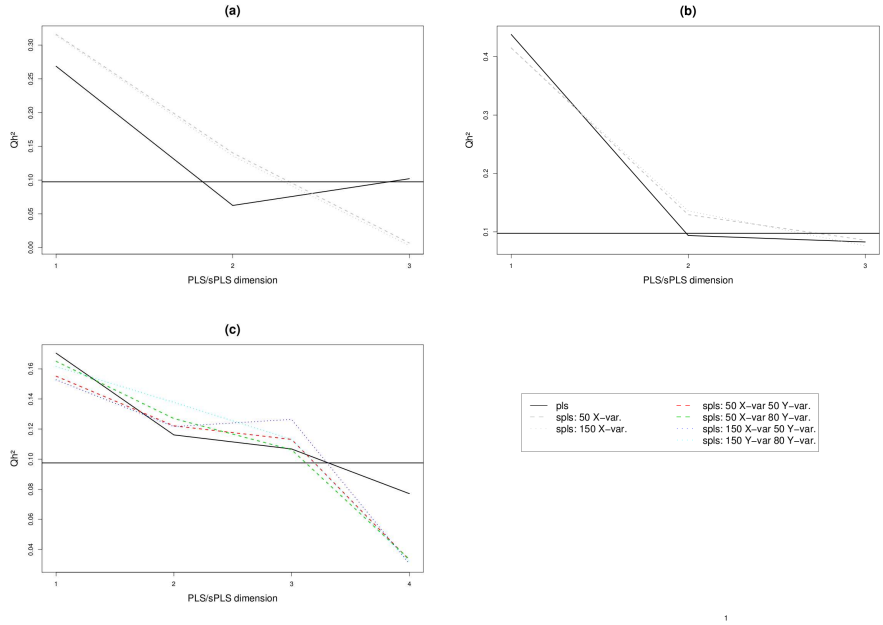


Figure 2: Marginal contribution of the latent variable ξ_h for each component in PLS and sPLS for different sparsity degrees for Liver Toxicity Study **(a)**, Wine Yeast **(b)** and Arabidopsis **(c)**. The horizontal black line indicates the threshold value in Q_h^2.

cycle, cell fate, protein bio synthesis and ribosome bio genesis, leaving a set of 3391 genes for a regression framework analysis, with no missing data, and $n = 43$ samples.

2.2.2 Comparisons with PLS

Comparisons with PLS will be performed in terms of criteria defined in section 1.6: Q_h^2, predictive power assessment of the model as well as insight into the variable selection in terms of stability. As the main objective of this paper is to show the feasibility of the sparse approach, the three data sets will be used as illustrative examples to compare PLS and sPLS. In this regression framework, some of the data sets are characterized by a very small q (Liver Toxicity: $q = 10$, Wine Yeast $q = 22$). In these cases, we did not judge relevant to perform selection on these Y variables, and hence $\lambda_2^h = 0$. In the other data set Arabidopsis, the selection was simultaneously performed on the X and Y data sets, as initially proposed by our approach.

Each input matrix was centered to column mean zero, and scaled to unit variance so as to avoid any dominance of one of the two-block data sets. Missing values were imputed with the NIPALS algorithm.

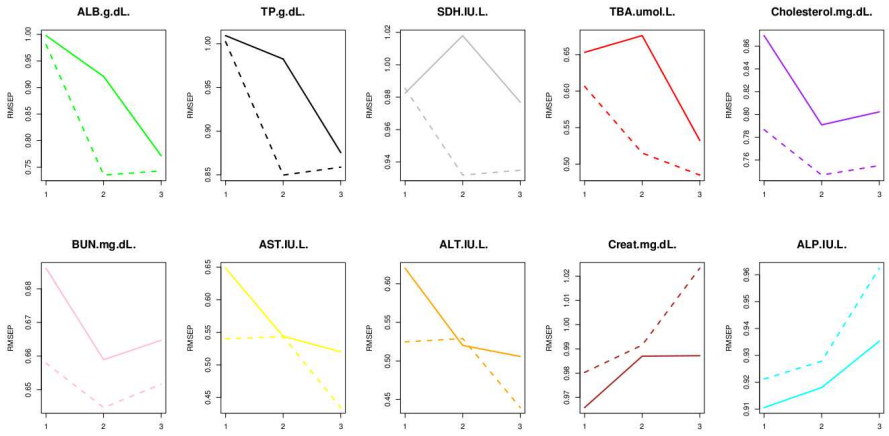


Figure 3: Liver Toxicity study: RMSEP for each clinical variable with PLS (plain line) and sPLS (dashed). Clinical variables are ranked according to their loadings in dimension 2.

Q^2_h. We compare the Q^2_h value with the PLS model with all variables in the model, and the proposed sparse PLS model with different sparsity degrees on each dimension : selection of 50 or 150 X variables on Liver Toxicity and Wine Yeast, 50 or 150 X variables coupled with the selection of 50 or 80 Y-variables in Arabidopsis. The choice of the selection size is arbitrarily chosen and loo-cv is applied for all data sets. The marginal contribution of ξ_h for each PLS/sPLS component is computed for each dimension. Figure 2 shows that the values of Q^2_h behave differently, depending on the data set and on the PLS/sPLS approach.

In Liver Toxicity and Wine Yeast **(a) (b)**, PLS needs one less component than sPLS : 1 (2) PLS dimensions for Liver Toxicity (Wine Yeast). As already observed in section 2.1.3, sPLS would need one more dimension to fully separate the different biological effects and select the X and Y variables according to each of these effects.

In Liver Toxicity, Q^2_3 increases and becomes superior to the threshold value 0.0975. On the other hand, the Q^2_h values in any sPLS steadily decreases with h.

In Arabidopsis **(c)** that is characterized by many X variables, and where a simultaneous variable selection is performed on the Y data set, the Q^2_h values differ depending on the number of variables that are selected on both data sets. However, for both methods and all sparsity degrees, the choice of $H = 3$ seems sufficient.

Predictive ability. Figure 3 compares the RMSEP for each clinical variable in the Liver Toxicity study with PLS (no selection) and sPLS (here, selection of 150 genes). As observed in section 2.1.2, these graphics show that except for 2 clinical variables, sPLS clearly outperforms PLS. Removing some of the noisy variables in the X data set helps for a better prediction of most of the Y variables. In this figure, the clinical

Table 3: Stability : ratio of the true positive variables selected in original data set and bootstrap data sets over the length of each selection (100).

	Liver toxicity		Arabidopsis				Wine Yeast	
	PLS	sPLS	PLS		sPLS		PLS	sPLS
			X	Y	X	Y		
dim 1	0.735	0.739	0.332	0.895	0.377	0.893	0.596	0.598
dim 2	0.457	0.603	0.221	0.834	0.365	0.838	0.622	0.559
dim 3	0.354	0.279	0.101	0.77	0.156	0.78	0.52	0.463

Table 4: Number of variables commonly selected in PLS (two step selection procedure) and in sPLS when selecting 100 variables.

	Liver toxicity	Arabidopsis		Wine Yeast
	X	X	Y	X
dim 1	97	56	90	91
dim 2	56	45	82	73
dim 3	19	72	80	74

variables are ranked according to the absolute value of their loadings in v_2. Hence the Y-loadings have a meaning in terms of variable importance measure, as the less better explained variables *creat.mg.dL* and *ALP.IU.L* get the lowest ranks. A thorough biological interpretation would be needed here to verify if these clinical variables are relevant in the biological study.

If the clinical variables were ranked according to the next loading v_3, then, although the graphics would be unchanged, *creat.mg.dL* and *ALP.IU.L* would get a a higher rank (resp. rank 1 and 8). This result comforts the choice of $H = 2$ for Liver Toxicity with sPLS. Similar conclusions can be drawn on the other data sets that includes more Y variables.

Variable selection.

Stability. On B bootstrap samples, $B = 10$, we compare the 100 X variables and 100 Y variables (in the case of Arabidopsis) that were selected either with PLS (two step selection procedure) or sPLS with respect to the same number of variables selected on the original data sets. The results are summarized in Table 3 and show that except for Wine Yeast in dimension 2 and 3, the sparse PLS approach seems more stable than PLS. It is not surprising to find an increased stability when the total number of variables (p and q) is rather small.

Comparison with PLS. Table 4 highlights the actual differences between a selection performed either with PLS (in two steps) or with sPLS for the same number of variables (100 for each data set, when applicable). As expected, both selections should be similar in dimension 1, but differ greatly for the other dimensions. In particular, the selections performed in the X Arabidopsis data set differ from the very first dimension. This is due to the extremely large number of X variables ($p = 22810$), where many of the transcripts get similar weights in PLS.

2.2.3 Property of the loading vectors.

When applying sparse methods, the loadings may lose their property of orthogonality and uncorrelation, as it was observed with sparse PCA (Trendafilov and Jolliffe, 2006; Shen and Huang, 2007). This is not the case with sPLS. In the original PLS, no constraint is set to have $\omega_r'\omega_s = 0$, $r < s$. Hence, latent variables $(\omega_1, \ldots, \omega_H)$ from the Y data set are not orthogonal in PLS or sPLS. To remedy to this in terms of graphical representation of the samples, we propose to re project $(\omega_1, \ldots, \omega_H)$ in an orthogonal basis. For the latent variables ξ, however, we always observed that $\xi_r'\xi_s = 0$ and no re projection is needed for these latent variables.

3 Analysis of the Wine Yeast data set and biological interpretation

We first give some elements of discussion regarding the graphical representation of the latent variables (samples), which facilitate the biological interpretation. These preliminary remarks will explain some of the results obtained when we compared the genes selected with PLS (two-step procedure) to the genes selected in the one step procedure with sPLS. Finally we show that the sPLS selection gives meaningful insight into the biological study.

As required by the biologists who performed this experiment, 200 genes were selected for each dimension.

3.1 Biological samples

Figure 4 highlights several facts that can actually be explained by the biological experiment. The plots of (ξ_1, ξ_2) (top) and (ω_1, ω_2) gave similar representation (not shown). The first component separated samples into time-specific clusters. This is to be expected as the particular stage of fermentation is the major source of genetic variation and the main determinant of aroma compound levels. The next most significant source of biological variation is the identity of the yeast strain. This was corroborated by the second and third components, where the samples clustered together in biological repeats of the same strain. Strains that are known to be more similar in terms of their

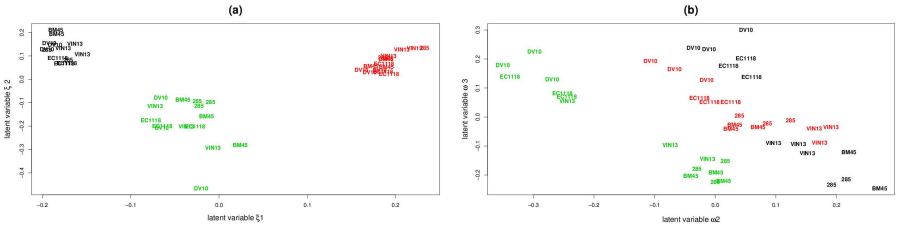


Figure 4: Wine Yeast data : graphical representation of the samples for the latent vectors (ξ_1, ξ_2) **(a)** and (ω_2, ω_3) **(b)**. Colors red, green and black stand for fermentation day 2, 5 and 14.

Table 5: Comparison of genes selected with PLS (two step procedure) vs. sPLS.

	PLS	sPLS
dim 1	-genes related to general central carbon metabolism -inclusion of many dubious/suspect ORFs	-GDH1: key regulator of cellular redox balance (direct influence on the main aroma producing reactions
dim 2	-identifies 'rate-limiting' enzymes in aroma metabolism	-improved coverage of transcriptional pathways
dim 3	-identifies most important alcohol and aldehydes dehyrogenase genes	-IDH1: key enzyme controlling flux distribution between aroma producing pathways and TCA cycle -NDE1: provides energy intermediates for dehydrogenase reactions

fermentative performance also clustered closely within time (*i.e* EC1118 and DV10, and BM45 and 285). The VIN13 strain (which is least similar to any of the other strains in this study) showed an intermediate distribution between the latent variable axes.

3.2 Selected variables

Comparisons with PLS Table 5 presents the similarities and main differences observed between the genes selected either with PLS or sPLS in regression mode. We adopted a two-step procedure to select genes with the original PLS approach by ordering the absolute values of the loadings u_h for each dimension ($H = 3$) and selecting the same number of top genes as in sPLS.

The striking result that we observed was the differences in the genes selections, especially in dimension 2 and 3. Overall, these dimensions were found to be more enriched for genes with proved or hypothesized roles in aroma compound production (based on pathway analysis and functional categorisation) for the sPLS rather than PLS.

Genes selected with sPLS. Figure 5 depicts the 'known' or hypothesised reactions and enzyme activities involved in the reaction network of higher alcohol and ester production. From the figure it is clear that the sPLS outputs provided good coverage

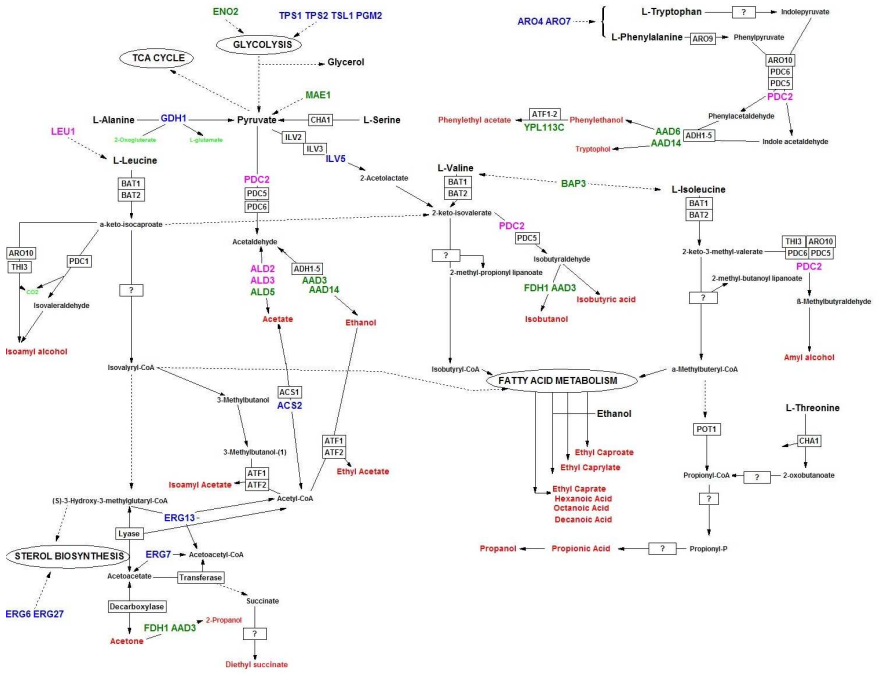


Figure 5: Graphical representation of 'known' or hypothesised reactions and enzyme activities involved in the reaction network of higher alcohol and ester production. Indirect interactions (*i.e* missing intermediates) are indicated by dashed lines and standard reactions are indicated by solid lines. Aroma compounds (red) and other metabolic intermediates (black) are positioned at the arrow apices. Unknown enzyme activities are represented by a question mark (**?**). Gene names coding for the relevant enzymes are represented in black box format, except for those genes that were identified in the first (blue), second (purple) and third (green) components of the sPLS.

of key reactions and major branches of the aroma production pathways (for the areas of metabolism with known reactions and enzymes). The first component identified mostly genes that are involved in reactions that produce the key substrates for starting points of the pathways of amino acid degradation and higher alcohol production. Amino acid metabolism is also a growth stage-specific factor (linked to fermentative stage), which is supported by the observations discussed in section 3.1. Most of the crucial 'rate limiting' enzymes (PDC2, ALD2, ALD3, LEU1) were identified by the second component. In total, the highest number of relevant genes were identified by the third component. Genes in this component were also interesting from the perspective that they only have putative (but unconfirmed) roles to play in the various pathways where they are indicated in the figure. Associations between genes with putative functional designations (based on homology or active site configuration) and aroma compounds in the lesser annotated branches of aroma compound production provide opportunities for directed research and the formulation of novel hypothesis in these areas.

Further analysis to be done. An attractive way of representing variables is to compute the correlation between the original data set (X and Y) and the latent variables $(\xi_1, \dots \xi_H)$ and $(\omega_1, \dots \omega_H)$, as it is done with PCA or CCA. These graphical representations, where the selected variables are projected on a correlation circle, will allow to identify known and unknown relationships between the X variables, the Y variables, and more importantly between both types of omics data. Of course these relationship will then need to be biologically assessed with further experiments, and will constitute a next step of our proposed analysis.

4 Conclusion

We have introduced a general computational methodology that modifies PLS, a well known approach that has been proved to be extremely efficient in many data where $n << p + q$, in a sparse version including variable selection to be more useful to the biologists. Validation of the sparse PLS approach has been performed both on simulated but also on real data sets and compared with PLS. The simulation study showed that sPLS selected the relevant variables from both data sets, that were governed by the known latent effects. The application to real data sets showed that this built-in variable selection procedure improved the predictive ability of the model, differed from PLS from dimension 2 and seemed more stable. Compared to PLS, sPLS seemed to highlight each latent biological effect on a different dimension and accordingly select the variables governed by each effect. This result will help biologists identifying relevant variables linked to each biological condition.

Our proposed algorithm is fast to compute. Like any sparse multivariate method, sPLS requires the addition of penalization parameters. The tuning of these two param-

eters can simply be performed by choosing the variable selection size, a useful option for the biologists. The gain by penalizing, and hence selecting variables, is proved on a typical biological study aiming at integrating gene expressions and metabolites in Wine Yeast. We provide a thorough biological interpretation and show that the sPLS results are extremely meaningful for the biologist, compared to a PLS selection. This preliminary work undoubtedly brought more insight into the biological study and will suggest further experiments to be performed.

Integrating omics data is an issue that may soon be commonly encountered in most high throughput biological studies. Hence we believe that our sparse PLS provides an extremely useful tool for the biologist in need of integrating two-block data sets and easily interpreting the resulting variable selections.

Remark. Another variant in our sparse PLS approach can be considered in step (g) of the proposed algorithm in section 1.4, by deflating the Y matrix in a symmetric manner: $Y_h = Y_{h-1} - \omega_h e_h'$. In this case, we are in a canonical framework and the aim is to model a *reciprocal relationship* between the two sets of variables. The lack of statistical criteria in this setting (as we are not in a predictive context) would require a thorough biological validation of the approach, rather than a statistical validation, and will constitute the next step of our research work.

Availability The code sources of sparse PLS (in R[1]) can be available upon request to the corresponding author. An R package is currently being implemented.

Acknowledgement

We are very grateful to Yves Gibon (Max Planck Institute of Molecular Plant Physiology) who kindly provided the full Arabidopsis data set.

References

Bely, M., Sablayrolles, J., and Barre, P. (1990). Description of Alcoholic Fermentation Kinetics: Its Variability and Significance. American Journal of Enology and Viticulture, 41(4):319–324.

Boulesteix, A. (2004). PLS Dimension Reduction for Classification with Microarray Data. Statistical Applications in Genetics and Molecular Biology, 3(1):1075.

Boulesteix, A. and Strimmer, K. (2005). Predicting transcription factor activities from combined analysis of microarray and chip data: a partial least squares approach. Theor Biol Med Model, 2(23).

Burnham, A., Viveros, R., and Macgregor, J. (1996). Frameworks for latent variable multivariate regression. Journal of chemometrics, 10(1):31–45.

Bushel, P., Wolfinger, R. D., and Gibson, G. (2007). Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. BMC Systems Biology, 1(15).

Bylesjö, M., Eriksson, D., Kusano, M., Moritz, T., and Trygg, J. (2007). Data integration in plant biology: the o2pls method for combined modeling of transcript and metabolite data. The Plant Journal, 52:1181–1191.

[1]The Comprehensive R Archive Network, http://cran.r-project.org/

Chun, H. and Keles, S. (2007). Sparse Partial Least Squares Regression with an Application to Genome Scale Transcription Factor Analysis. Technical report, Department of Statistics, University of Wisconsin, Madison, USA.

Culhane, A., Perriere, G., and Higgins, D. (2003). Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. BMC Bioinformatics, 4(1):59.

Datta, S. (2001). Exploring relationships in gene expressions: A partial least squares approach. Gene Expr, 9(6):249–255.

de Jong, S. (1993). Simpls: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 18:251–263.

De Jong, S. and Ter Braak, C. (1994). Comments on the PLS kernel algorithm. Journal of chemometrics, 8(2):169–174.

Dickinson, J., Salgado, L., and Hewlins, M. (2003). The Catabolism of Amino Acids to Long Chain and Complex Alcohols in Saccharomyces cerevisiae. Journal of Biological Chemistry, 278(10):8028–8034.

Doledec, S. and Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biology, 31(3):277–294.

Dray, S., Pettorelli, N., and Chessel, D. (2003). Multivariate Analysis of Incomplete Mapped Data. Transactions in GIS, 7(3):411–422.

Gibon, Y., Usadel, B., Blaesing, O., Kamlage, B., Hoehne, M., Trethewey, R., and Stitt, M. (2006). Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in Arabidopsis rosettes. Genome Biology, 7:R76.

Gidskehaug, L., Anderssen, E., Flatberg, A., and Alsberg, B. (2007). A framework for significance analysis of gene expression data using dimension reduction methods. BMC Bioinformatics, 8(1):346.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 46(1):389–422.

Heinloth, A., Irwin, R., Boorman, G., Nettesheim, P., Fannin, R., Sieber, S., Snell, M., Tucker, C., Li, L., Travlos, G., et al. (2004). Gene Expression Profiling of Rat Livers Reveals Indicators of Potential Adverse Effects. Toxicological Sciences, 80(1):193–202.

Hoskuldsson, A. (1988). PLS regression methods. Journal of Chemometrics, 2(3):211–228.

Lê Cao, K.-A., Gonçalves, O., Besse, P., and Gadat, S. (2007). Selection of biologically relevant genes with a wrapper stochastic algorithm. Statistical Applications in Genetics and Molecular Biology, 6(:Iss. 1):Article 1.

Lorber, A., Wangen, L., and Kowalski, B. (1987). A theoretical foundation for the PLS algorithm. Journal of Chemometrics, 1(19-31):13.

Mevik, B.-H. and Wehrens, R. (2007). The pls package: Principal component and partial least squares regression in r. Journal of Statistical Software, 18(2).

Nykanen, L. and Nykanen, I. (1977). Production of esters by different yeast strains in sugar fermentations. J. Inst. Brew, 83:30–31.

Pihur, V., Datta, S., and Datta, S. (2008). Reconstruction of genetic association networks from microarray data: A partial least squares approach. Bioinformatics.

Ribéreau-Gayon, P., Dubourdieu, D., Donche, B., and Lonvaud, A. (2000). Biochemistry of alcoholic fermentation and metabolic pathways of wine yeasts in Hand volume 1. John Wiley and Sons.

Shen, H. and Huang, J. Z. (2007). Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis, to appear.

Tenenhaus, M. (1998). La régression PLS: théorie et pratique. Editions Technip.

Trendafilov, N. and Jolliffe, I. (2006). Projected gradient approach to the numerical solution of the SCoTLASS. Computational Statistics and Data Analysis, 50(1):242–253.

Trygg, J. and Wold, S. (2003). O2-pls, a two- block (x-y) latent variable regression (lvr) method with an integral osc filter. Journal of Chemometrics, 17:53–64.

Umetri, A. (1996). SIMCA-P for windows, Graphical Software for Multivariate Process Modeling. Umea, Sweden.

Waaijenborg, S., de Witt Hamer, V., Philip, C., and Zwinderman, A. (2008). Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis. Statistical Applications in Genetics and Molecular Biology, 7(1):3.

Wegelin, J. (2000). A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Technical Report 371, Department of Statistics, University of Washington, Seattle.

Wold, H. (1966). Multivariate Analysis. Academic Press, New York, Wiley, krishnaiah, p.r. (ed.) edition.

Wold, S., Eriksson, L., Trygg, J., and Kettaneh, N. (2004). The PLS method–partial least squares projections to latent structures–and its applications in industrial RDP (research, development, and production). Technical report, Umea University.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B, 67(2):301–320.

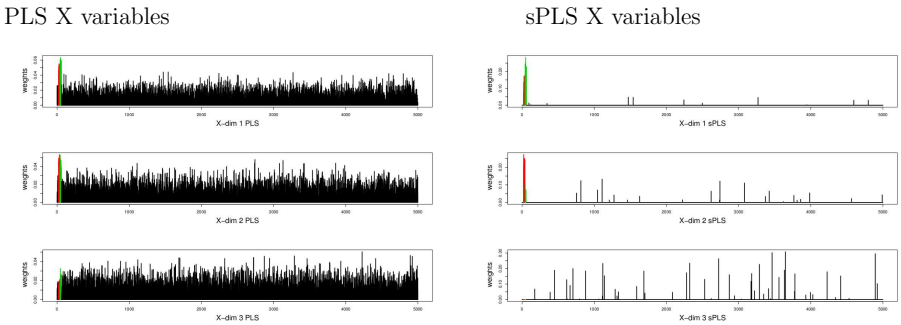


Figure 6: Supplemental figure: absolute variable weights in the loading vectors of PLS (left) or sparse PLS (right) for the 5000 X variables. Red (green) color stands for the variables related to the μ_1 (μ_2) effect.

9. Article appliqué

Cet article est une étude comparative de plusieurs approches "sparse" dans une étude biologique nécessitant un contexte canonique. Trois approches sont comparées sur un même jeu de données : l'analyse de Co-Inertie (CIA, Doledec & Chessel 1994, avec sélection de variables en deux temps), l'analyse Canonique des Corrélation avec pénalisation Elastic Net (CCA-EN, Waaijenborg *et al.* 2008) et notre approche sparse PLS mode canonique (sPLS). Ces deux dernières approches incluent la sélection de variables en une seule étape. Nous choisissons de baser l'étude comparative principalement sur des critères biologiques, pour comparer les différents objectifs de chaque approche, et pour convaincre le biologiste de l'utilité de ces méthodes.
Nous montrons que biologiquement, les approches CCA-EN et sPLS apportent des résultats similaires et pertinents, tandis que CIA semble apporter de l'information redondante dans les listes des variables sélectionnées. L'interprétation biologique des résultats a fait l'objet d'une étude approfondie grâce au logiciel *Ingenuity Pathways Analysis*.

Cet étude fera l'objet d'un article qui sera très prochainement soumis.

Sparse canonical methods for biological data integration: application to a cross-platform study

Kim-Anh Lê Cao[1,2,4*] and Pascal G.P. Martin,[3,4] Christèle Robert-Granié [2], Philippe Besse [1]

September 13, 2008

Abstract

In the context of integration for systems biology, very few sparse approaches have been proposed so far to select variables in a canonical framework. In this study we propose a canonical mode of a new sparse PLS approach to handle two-block data sets, where the relationship between the two types of variables is known to be symmetric. Sparse PLS has been proposed for either a regression or a canonical mode and includes a built-in procedure to perform variable selection while integrating data. To illustrate the canonical mode approach, we analyzed the NCI60 data sets, where two different platforms (cDNA and Affymetrix chips) were used to study the transcriptome of sixty cancer cell lines.
We compare the results obtained with two other sparse or related canonical approaches: CCA with Elastic Net penalization (CCA-EN) and Co-Inertia Analysis (CIA). The latter does not include a built-in procedure for variable selection and requires a two-step analysis. We stress the lack of statistical criteria to evaluate canonical methods, which makes biological interpretation crucial to compare the different gene lists. We propose comprehensive

*to whom correspondence should be addressed: Kim-Anh.Le-Cao@toulouse.inra.fr
[1]Institut de Mathématiques, Université de Toulouse et CNRS (UMR 5219), F-31062 Toulouse, France
[2]Station d'Amélioration Génétique des Animaux UR 631, Institut National de la Recherche Agronomique, F-31326 Castanet, France
[3]Laboratoire de Pharmacologie et Toxicologie UR 66, Institut National de la Recherche Agronomique, F-31931 Toulouse, France
[4]Both authors contributed equally to this work.

graphical representations of both samples and variables to facilitate the biologist interpretation.
We show that sPLS and CCA-EN select highly relevant genes, which enable a detailed understanding of the molecular characteristics of several groups of cell lines. These two approaches were found to bring similar results, although they highlighted the same phenomenons with a different priority. On the other hand, CIA tended to select redundant information. These canonical methods seem to be efficient tools to deal with variable selection in the context of high-throughput data integration.

Introduction

When dealing with the integration of high dimensional biological data, the application of linear multivariate models such as Partial Least Squares regression (PLS, Wold, 1966) and Canonical Correlation Analysis (CCA, Hotelling, 1936), are often limited by the size of the data set (ill-posed problems), the noisy and the multicollinearity characteristics of the data and the lack of interpretability (PLS). However, these approaches still remain extremely interesting for this type of problems, first because they allow for the compression of the data into 2 to 3 dimensions for a more powerful and global view, and second as their resulting components and loading vectors capture dominant and latent properties of the studied process. They may hence provide a better understanding of the underlying biological systems, for example by revealing groups of samples that were previously unknown or uncertain.
In this study, we were interested in integrating two high dimensional data sets, where variables of two types are measured on the same individuals or samples. Recent integrative biological studies applied Principal Component Analysis, or PLS (Bylesjö et al., 2007; Vijayendran et al., 2008), but in a regression framework, where prior biological knowledge indicates which type of omic data is expected to explain the other type (for example transcripts and metabolites).
Here, we specifically focus on a canonical framework, when there is either no assumption on the relationship between the two sets of variables (exploratory approach), or when a reciprocal relationship between the two sets

is expected (*e.g.* cross platform comparisons).
Few statistical methods can answer this problem. Among them, some are limited by the number of variables (CCA) or do not give straightforward interpretable results when the number of variables is too large (PLS). Some associated approaches have recently been developed to include a built-in selection procedure, so as to allow variable selection in both data sets. These sparse methods adapt lasso penalty (Tibshirani, 1996) or combine lasso and ridge penalties (Elastic Net, Zou and Hastie, 2005) for feature selection in integration studies.
In this study, we propose a sparse canonical approach called "sparse PLS" (sPLS) in the context of integration for systems biology. Methodological aspects and analyses of the sPLS in a regression framework were presented in (Lê Cao et al., 2008). This novel computational method provides variable selection of two-block data set in a one step procedure, for integrating variables of two types.
When applying canonical methods, most validation criteria used in a regression context are not statistically meaningful. Instead, the biological relevancy of the results should be evaluated during the validation process. We therefore compare sparse PLS with two other canonical approaches: penalized CCA adapted with Elastic Net (Waaijenborg et al., 2008), which is a sparse method that was applied to relate gene expression with gene copy numbers in human gliomas, and Co-Inertia Analysis (CIA, Doledec and Chessel, 1994) that was first developed for ecological data, and then for canonical high-throughput biological studies (Culhane et al., 2003). This latter approach does not include feature selection, which has to be performed in a two-step procedure.
This comparative study has two aims. First to better understand the main differences between each of these approaches and identify which method would be appropriate depending on the biological question. Second to highlight how each method is able to reveal the underlying biological processes inherent to the data. This type of comparative analysis renders the biological interpretation mandatory to strengthen the statistical hypothesis, especially when there is a lack of statistical criteria to assess the validity of the results.

We first recall some canonical methods among which the two sparse methods will be compared with CIA on the NCI60 cell lines data set, which is fully described. We propose to use appropriate graphical representations to discuss the results. The different gene lists are assessed, first with some statistical criteria, and then through their biological interpretation. Finally we discuss the pros and cons of each tested approach before concluding.

1 Canonical Methods

We focus on two-block data matrices denoted $X(n \times p)$ and $Y(n \times q)$, where the p variables x^j and q variables y^k are two types of measures performed on the same samples or individuals, $j = 1 \dots p$, $k = 1 \dots q$. Prior biological knowledge on these data allow us to settle into a canonical framework, *i.e.* there exists a reciprocal relationship between the X variables and the Y variables. In the case of high throughput biological data, the large number of variables may affect the exploratory method, due to numerical issues (as it is the case for example with CCA), or lack of interpretability (PLS).
We next recall three types of multivariate methods (CCA, PLS, CIA). For CCA and PLS, we recall their associated sparse approaches that were proposed, either to select variables from each set or to deal with the ill-posed problem commonly encountered in high-throughput biological data.

1.1 Canonical Correlation Analysis

Canonical Correlation Analysis (Hotelling, 1936) studies the relationship between two sets of data. The CCA n-dimensional score vectors (Xa_h, Yb_h) come in pairs to solve the objective function:

$$\arg \max_{a_h' a_h = 1, b_h' b_h = 1} cor(Xa_h, Yb_h), \qquad h = 1 \dots H$$

where the p- and q-dimensional vectors a_h and b_h are called canonical factors, or loading vectors, and h is the CCA chosen dimension.
As $cor(Xa_h, Yb_h) = cov(Xa_h, Yb_h)/\sqrt{var(Xa_h)}\sqrt{var(Yb_h)}$, the aim of CCA is to simultaneously maximize $cov(Xa_h, Yb_h)$ and minimize the variances of

Xa_h and Yb_h.
In the $p+q >> n$ framework, CCA suffers from the high dimensionality as it requires the computation of two inverses of the covariance matrices XX' and YY' that are singular. This implies numerical difficulties, since the canonical correlation coefficients are not uniquely defined. One solution proposed by Vinod (1976) was to introduce l_2 penalties in a ridge CCA (rCCA) on the covariance matrices, so as to make them invertible. González et al. (2008b) applied rCCA to post genomic data (Combes et al., 2008) and proposed to choose the optimal penalization parameters with cross-validation.
It is known (Gittins, 1985) that the CCA loadings are not directly interpretable. It is however very instructive to interpret these components by calculating the correlation between the original data set X and $\{a_1, \ldots, a_H\}$ and similarly between Y and $\{b_1, \ldots, b_H\}$, to project the variables on correlation circles. Easier interpretable graphics are obtained, which readability was improved by González et al. (2008b) in the `R` package `cca`. In our study, rCCA could not be applied as it does not perform feature selection. Furthermore, because of the non direct interpretability of the loadings, a variable selection in a two-step procedure is difficult to perform, as it must be based on correlation circles graphics.

1.2 PLS

Partial Least Squares regression (Wold, 1966) is based on the simultaneous decomposition of X and Y into latent variables and associated loading vectors. The latent variables methods (*e.g.* PLS, Principal Component Regression) assume that the studied system is driven by a small number of n-dimensional vectors called latent variables. These latter may correspond to some biological underlying phenomenons which are related to the study (Wold et al., 2004).
Like CCA, the PLS components (latent variables) are linear combinations of the predictor variables, but the objective function differs as it is based on the maximization of the covariance between each linear combination of the two sets of variables:

$$\arg \max_{a_h' a_h = 1, b_h' b_h = 1} cov(Xa_h, Yb_h), \qquad h = 1 \ldots H.$$

We denote $\xi_h = Xa_h$ and $\omega_h = Yb_h$ the latent variables associated to each loading vector a_h and b_h, h being the chosen PLS dimension. On one hand, and in contrary to CCA, the loading vectors (ξ_h, ω_h) are interpretable and can give information about how the x^j and y^k variables combine to explain the relationships between X and Y. On the other hand, the PLS latent variables (a_h, b_h) indicate the similarities or dissimilarities between the individuals, related to the loading vectors.
Many PLS algorithms exist, not only for different shapes of data (SIMPLS, de Jong, 1993, PLS1 and PLS2, Wold, 1966, PLS-SVD, Lorber et al., 1987) but also for different aims (predictive, like PLS2 or modelling, like PLS-mode A, see Tenenhaus, 1998; Wegelin, 2000; Waaijenborg et al., 2008). In this study we especially focus on a modelling aim (canonical mode) between the two data sets, by deflating X and Y in a symmetric way (see appendix A).

1.3 Penalized Correlation Analysis with Elastic Net (CCA-EN)

Waaijenborg et al. (2008) proposed a sparse penalized variant of CCA using Elastic Net (Zou and Hastie, 2005; Zou et al., 2006) in a regression framework. To do so, the authors used the PLS-mode A formulation (Tenenhaus, 1998; Wegelin, 2000) to introduce penalties. Note that Elastic Net is well adapted in this particular framework. It combines the advantages of the ridge regression, that penalizes the covariance matrices XX' and YY' which become non singular, and the lasso (Tibshirani, 1996) that allows variable selection, in a one step procedure. However, when $p + q$ is very large, the resolution of the optimization problem requires intensive computations, and Zou and Hastie (2005); Waaijenborg et al. (2008) proposed instead to perform a univariate thresholding, that leaves only the lasso estimates to compute (see appendix C).

1.4 sparse PLS

Lê Cao et al. (2008) proposed a sparse PLS approach (sPLS) based on a PLS-SVD variant, so as to penalize both loading vectors a_h and b_h simultaneously.

For any matrix M($p \times q$) of rank r, the SVD of M is given by:

$$M = A\Delta B'$$

where the columns of A ($n \times r$) and $B(r \times p)$ are orthonormal and contain the eigenvectors of MM' and $M'M$, Δ is a diagonal matrix of the squared eigenvalues of MM' or $M'M$.
If $M = X'Y$, then the column vectors of A (resp. B) correspond to the loading vectors of the PLS, and sparsity in both vectors can be introduced by iteratively penalizing a_h and b_h with a soft-thresholding penalization, as was proposed in a similar manner by Shen and Huang (2007) for a sparse PCA (see appendix B for more details). Both deflation modes, as referred in section 1.2, were proposed. In this paper, we will focus on the canonical mode only. The regression mode has been already been discussed in Lê Cao et al. (2008) where a thorough biological interpretation was provided in this framework.

1.5 Co-Inertia Analysis

Co-Inertia analysis (CIA) was first introduced by Doledec and Chessel (1994) in the context of ecological data, before Culhane et al. (2003) applied it to high-throughput biological data. CIA is suitable for a canonical framework, as it is adapted for a symmetric analysis. It involves analyzing each data set separately either with principal component analyses, or with correspondence analyses, such that the covariance between the two new sets of projected scores vectors (that maximize either the projected variability or inertia) is maximal. This results in two sets of axes, where the first pair of axes are maximally co-variant, and are orthogonal to the next pair (Robert and Escoufier, 1976).
CIA does not propose a built-in variable selection, but we can perform instead a two-step procedure by ordering the weight vector (loadings) for each CIA dimension and select the top variables.

1.6 Differences between the approaches

The three canonical approaches that we want to compare (CCA-EN, sPLS, CIA) profoundly differ in their construction, and hence their aims.

CCA-EN looks for canonical variate pairs (Xa_h, Yb_h), such that a penalized version of the canonical correlation is maximized. This explains why a non monotonic decreasing trend in the canonical correlation can sometimes be obtained (Waaijenborg et al., 2008). On the other hand, sPLS (canonical mode) and CIA aim at maximizing the covariance between the scores vectors, so that there is a strong symmetric relationship between both sets. However, here CIA is based on the construction of two Correspondence Analyses, whereas sPLS is based on a PLS analysis.

1.7 Parameters tuning

In CCA-EN, the authors proposed to tune the penalty parameters for each dimension, such that the canonical correlation $cor(Xa_h, Yb_h)$ is maximized. In practice, they showed that the correlation did not change much when variables were added in the selection. Hence, an appropriate way of tuning the parameters would be to choose instead the degree of sparsity (*i.e.* the number of variables to select), as proposed by Zou et al. (2006) for their sparse PCA in the `elasticnet` `R` package, and rely on the biologists needs. Indeed, a too short gene selection may lack in information, as some of the functions or annotations may be missing. The same strategy will be used for sPLS. No other parameters than the number of selected variables is needed in CIA either.

1.8 Outputs

Graphical representations should be a an important issue to help biologists interpret the results. Hence we propose to homogenize all ouputs to get comparable results.
Samples will be represented with the scores or latent vectors, in a superimposed manner, as proposed in the `R` package `ade4` (Thioulouse et al., 1997), first to show how samples are clustered based on their biological characteristics, and second to measure if both data sets strongly agree according to the applied approach.
Variables will be represented on correlation circles, as proposed by González et al. (2008b). Correlations between the original data sets and the loading

vectors are computed so that highly correlated variables will cluster together in the resulting graphics. Only the selected variables in each dimension will be represented. This type of graphic will allow to identify interactions between the two types of variables and relate the variable clusters to their associated sample clusters.

2 Cross-platform study

2.1 Data sets and relevance for a canonical analysis

We compared the three canonical methods (CCA-EN, CIA and sPLS) for their ability to highlight the relationships between two gene expression data sets both obtained on a panel of 60 cell lines (NCI60) from the National Cancer Institute (NCI). This panel consists of human tumor cell lines derived from patients with leukemias (LE), melanomas (ME) and cancers of ovarian (OV), breast (BR), prostate (PR), lung (LU), renal (RE), colon (CO) and central nervous system (CNS) origin. The NCI60 is used by the Developmental Therapeutics Program (DTP) of the NCI to screen thousands of chemical compounds for growth inhibition activity and it has been extensively characterized at the DNA, mRNA, protein and functional levels. The data sets considered here have been generated using Affymetrix (Butte et al., 2000; Staunton et al., 2001) or spotted cDNA (Ross et al., 2000) platforms. These data sets are highly relevant to an analysis in a canonical framework since 1) there is some degree of overlap between the genes measured by the two platforms but also a large degree of complementarity through the screening of gene sets representing common pathways or biological functions (Culhane et al., 2003) and 2) they play fully symmetric roles as one data set cannot be explained by the other, it as would be done in a regression framework. Considering that the aim of the canonical methods is to capture the relationships between two data sets, each of which should be relevant to the problem under study (here, the characteristics of the gene expression profiles of tumor cell lines of different origins), we believe that these methods should primarily apply to pre-processed data sets, where data transformation, background correction and normalization steps were performed beforehand. These steps and the resulting data sets that

were analyzed here are briefly described below.

2.2 The Ross Data set

Ross et al. (2000) used spotted cDNA microarrays containing 9,703 human cDNAs to profile each of the 60 cell line in the NCI60 panel (Ross et al., 2000). Here, we used a subset of 1,375 genes that has been selected using both non-specific and specific filters described in Scherf et al. (2000). In particular, genes with more than 15% of missing values were removed and the remaining missing values were imputed by k-nearest neighbours (Culhane et al., 2003). The pre-processed data set containing log ratio values is available in Culhane et al. (2003).

2.3 The Staunton Data set

Hu6800 Affymetrix microarrays containing 7,129 probe sets were used to screen each of the 60 cell lines in another study (Butte et al., 2000; Staunton et al., 2001). Pre-processing steps are described in Staunton et al. (2001) and Culhane et al. (2003). They include 1) replacing average difference values less than 100 by an expression value of 100, 2) eliminating genes whose expression was invariant across all 60 cell lines and 3) selecting the subset of genes displaying a minimum change in expression across all 60 cell lines of at least 500 average difference units. The final analyzed data set contained the average difference values for 1,517 probe sets, and is available in Culhane et al. (2003).

2.4 Application of the three canonical methods

We applied CCA-EN, CIA and sPLS to the Ross (X) and Staunton (Y) data sets. For each dimension h, $h = 1 \dots 3$, we performed variable selection of 100 genes from each data set. The number of dimensions was arbitrarily chosen, given that if $H \geq 4$, the interpretation of the results becomes more difficult due to the high number of graphical outputs, and the results were less relevant. The size of the selection (100) was judged small enough to allow for the identification of into individual relevant genes and large enough to reveal gene groups belonging to the same functional category or pathway.

The graphical representation of the individuals, as described in section 1.8, is displayed in a superimposed manner, where each sample will be indicated using an arrow. The start of the arrow will indicate the location of the sample in X in one plot, and the tip the location of the sample in Y in the other plot. Short arrows will therefore indicate if both data sets strongly agree and long arrows a disagreement between the two data sets.

3 Results and Discussion

We apply the three canonical approaches to the NCI60 data set and assess the results in two different ways. First we examine few statistical criteria, then we provide an interpretation of the results from each method, using graphical representations along with database mining.

3.1 How to assess the results ?

Canonical methods are statistically difficult to assess. Firstly because they do not fit into a regression/prediction framework, meaning that cross-validation cannot be computed to evaluate the quality of the model. Secondly because in many two-block biological studies, the number of samples n is very small compared to the number of variables $p + q$. This makes any statistical criteria difficult to compute. This is why graphical outputs are important to analyse the results (see for example Tenenhaus, 1998; Culhane et al., 2003).
When working with biological data, a new way of assessing the results should be to strongly rely on the biological interpretation. Indeed our aim is to show the applicability of each approach and to show if they answer the biological question. We hence propose to base most of our comparative study on the biological interpretation by using appropriate graphical representations of the samples and of the selected variables.

3.2 Link between two-block data set

Variance explained by each component. Tenenhaus (1998) proposed to estimate the variance explained in each data set X and Y in relation to the "op-

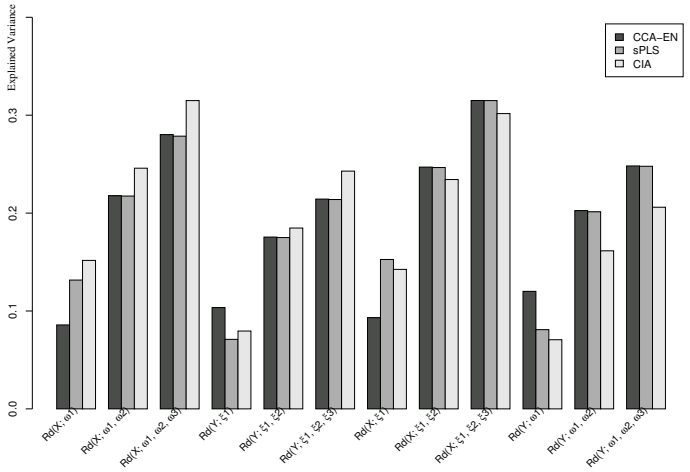


Figure 1: Cumulative explained variance of each data set in relation to its component score (CCA-EN, CIA) or latent variable (sPLS)

posite" component score or latent variables $(\omega_1, \ldots, \omega_H)$ and $(\xi_1, \ldots, \xi_H)$, where $\xi_h = Xa_h$ and $\omega_h = Yb_h$ in all approaches. The redundancy criterion Rd, or part of explained variance, is computed as follows:

$$Rd(X; \omega_1, \ldots, \omega_H) = \frac{1}{p} \sum_{h=1}^{H} \sum_{j=1}^{p} cor^2(x^j, \omega_h)$$

$$Rd(Y; \xi_1, \ldots, \xi_H) = \frac{1}{q} \sum_{h=1}^{H} \sum_{k=1}^{q} cor^2(y^k, \xi_h)$$

Similarly, one can compute the variance explained in each component in relation with its associated data set:

$$Rd(X; \xi_1, \ldots, \xi_H) = \frac{1}{p} \sum_{h=1}^{H} \sum_{j=1}^{p} cor^2(x^j, \xi_h)$$

$$Rd(Y; \omega_1, \ldots, \omega_H) = \frac{1}{q} \sum_{h=1}^{H} \sum_{k=1}^{q} cor^2(y^k, \omega_h)$$

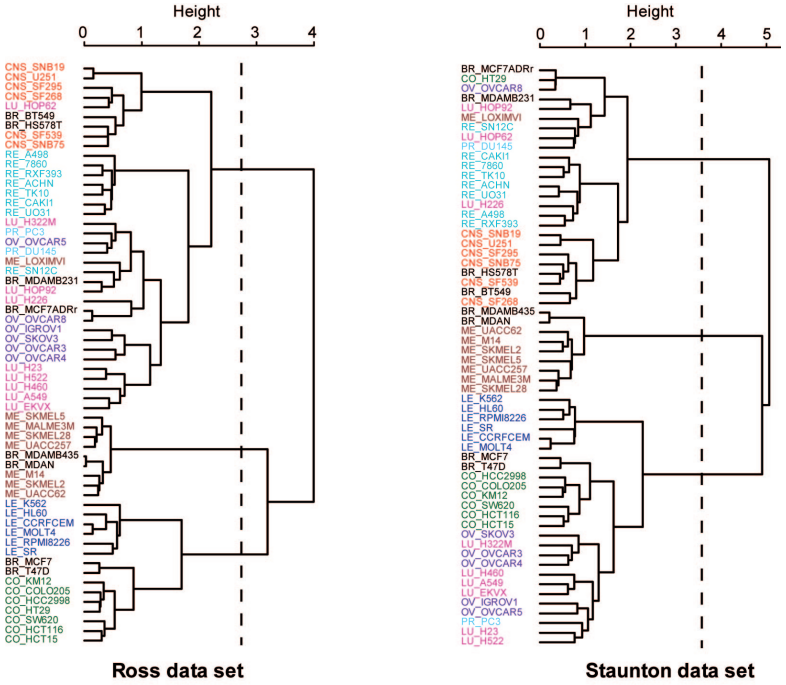


Figure 2: Hierarchical clustering of Ross and Staunton data sets expression profiles of the cell lines, which are coded as CO = Colon, ME = Melanoma, BR = Breast, CNS= Central Nervous System , OV= Ovarian, RE = Renal, PR = Prostate.

Figure 1 displays the *Rd* criterion for $h = 1 \dots 3$ for each set of components $(\xi_1, \dots, \xi)$, $(\omega_1, \dots, \omega_H)$ and for each approach. While there seems to be a great difference in the first dimension between CCA and the other methods, the components in dimensions 2 and 3 explain the same amount of variance in both X and Y for CCA-EN and sPLS. This suggests a strong similarity at this stage between these two approaches. On the other hand, CIA differs from these two methods. The components computed from the "opposite" set explain more variance than CCA/sPLS, and less in their respective set.
More generally, we can observe that more information seems to be present in the X rather than in the Y data set. Indeed, similarly to Culhane et al. (2003), we noticed that a hierarchical clustering of the samples using the distance $1-$correlation with the Ross data set allows a better clustering of the cell lines based on their tissue of origin than with the Staunton data set (Figure 2).

Correlations between each component. The canonical correlations between the pair of score vectors (or latent variables) were very high (>0.93) for any approach and in any dimension (see Table 1). This comfort our hypothesis regarding the canonical aim of each method.
The non monotonic decreasing trend in the canonical correlations in CCA-EN is not what can be expected from a CCA variant, but was also pointed out by Waaijenborg et al. (2008) as the optimization criterion differs from ordinary CCA. However, the computations of the *Rd* criterion (Figure 1) seem to indicate that the cumulative variance explained by the latent variables increases with h. In sPLS and CIA, which aim is to maximize the covariance, we can see that in fact they also highlight very strongly correlated components. This suggests that the associated loading vectors may also bring related information from both data sets.
The maximal canonical correlation ($\simeq 0.97$) is obtained on the first dimension for CCA-EN, and surprisingly only on the second dimension for CIA and sPLS. In the next sections, we will see that in fact CCA-EN and sPLS "swap" their components between the first and second dimensions.

Table 1: Correlations of the score vectors/latent variables for each dimension.

	CCA-EN	CIA	sPLS
$cor(\xi_1, \omega_1)$	0.967	0.935	0.938
$cor(\xi_2, \omega_2)$	0.937	0.967	0.964
$cor(\xi_3, \omega_3)$	0.953	0.955	0.944

3.3 Interpretation of the observed cell line clusters

Figures 3 and 4 display the graphical representations of the samples in dimension 1 and 2 **(a)**, or 1 and 3 **(b)** for CCA-EN (Fig. 3) and sPLS (Fig. 4), CIA showing patterns similar to sPLS and to those presented in Culhane et al. (2003).
All graphics show that both data sets are strongly related (short arrows), but depending on the applied approach, the components differ. In dimension 1, the pair (ξ_1, ω_1) tends to separate the melanoma cell lines from the other cell lines in CCA-EN (Fig. 3 **(a)**), whereas sPLS and CIA tend to separate the LE and CO cell lines on one side from the RE and CNS cell lines on the other side (Fig. 4 **(a)**). As previously proposed (Culhane et al., 2003), we interpreted this clustering of the cell lines along the first axes of sPLS and CIA as the separation of cell lines with *epithelial* characteristics (mainly LE and CO) from those with *mesenchymal* characteristics (in particular RE and CNS). Epihelial cell generally form layers by making junctions between them and interacting with the extracellular matrix (ECM). On the other hand, mesenchymal cells are able to migrate through the ECM and are found in the connective tissues. We will see that the interpretation of the genes lists selected on the axes separating LE and CO versus RE and CNS strongly argue for such an interpretation of the individuals plot. In addition, it has been previously described that glioblastoma cell lines (CNS) do express mesenchymal stem-like properties at multiple levels, including gene expression (Tso et al., 2006). In dimension 2, we observe the opposite tendency: the pair (ξ_2, ω_2) separates the cell lines with epithelial characteristics from the cell lines with mesenchymal characteristics in CCA-EN while it separates the melanoma samples from the other samples in sPLS and CIA.

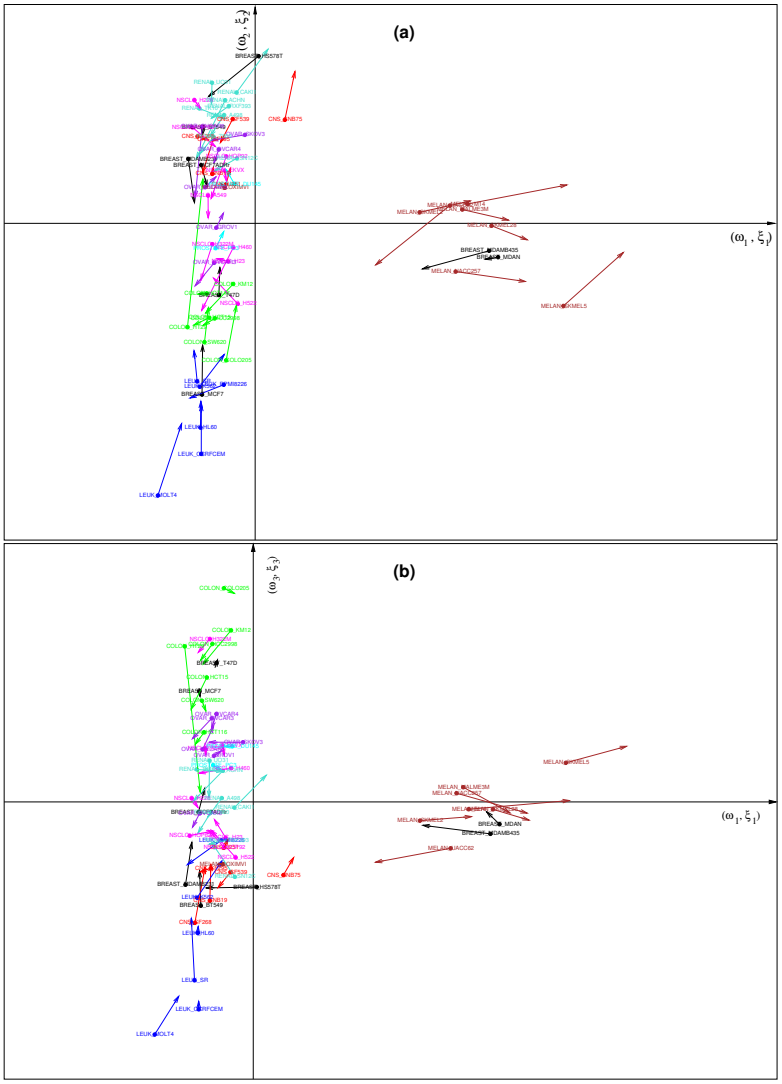


Figure 3: Graphical representations of the cell lines. CCA-EN component scores are displayed in a superimposed manner, where the start of the arrow show the location of the Ross samples, and the tip the Staunton samples. The first and second axis (first and third) are shown in **(a)** and **(b)**.

When performing hierarchical clustering of the 60 cell lines (with 1 −correlation distance) separately on each data set (Figure 2), it appears that the three main clusters of samples largely correspond to the three groups that are separated by all three methods in dimensions 1 and 2 *i.e.* they correspond to 1) cell lines with epithelial characteristics (LE and CO for both data sets), 2) cell lines with mesenchymal characteristics (in particular RE and CNS) and 3) melanomas with which two breast cancer cell lines (MDA_N and MDA_MB435) are systematically clustered. Among these clusters, only the third one is strictly identical for the two data sets. This illustrates that CCA-EN primarily captures the sample characteristics in the clusters that are most conserved between the two data sets, even if these do not underlie the separation of the main clusters within each data set. The fact that, based on their gene expression profiles, ME samples form a relatively homogeneous and compact cluster along with two breast tumor cell lines (MDA_N and MDA_MB435 which are indeed melanoma metastases derived from a patient diagnosed with breast cancer), has been previously shown by other authors (Ross et al., 2000; Scherf et al., 2000; Culhane et al., 2003) and seems largely independent of the initial gene selections that were used here. We believe that the strongest canonical correlation can only be found when separating this specific set of cell lines (see Table 1). This explains why CCA-EN, that looks for maximal correlation, first focuses on this particular axis. On the other hand, sPLS and CIA first focus on the separation between cell lines with epithelial versus mesenchymal characteristics, a separation that is slightly more obvious in the dendrograms obtained from the two data sets, but where the cluster members substantially change between the two data sets. In particular, most OV and LU cell lines are clustered with LE and CO in the Staunton data set while they are clustered with RE and CNS cell lines in the Ross data set (Figure 2). To further evaluate this hypothesis, we permuted the labels from 1 to 4 (out of 7) melanoma cell lines with randomly selected cell lines in one of the data set, thus artificially reducing the consistency between the clustering of the melanoma cell lines in the two data sets. Resulting graphics in CCA-EN happened to be similar to those obtained for sPLS and CIA in the absence of permutation (Figure 3 **(a)**), hence separating epithelial-

like versus mesenchymal-like cell lines on the first dimension. By contrast, sPLS and CIA graphics remained the same after the permutations.

3.4 Interpretation of the observed genes clusters

We computed the correlations between the original data sets and the scores vectors or latent variables (ξ_1, ξ_2, ξ_3) and $(\omega_1, \omega_2, \omega_3)$. Only the genes selected in each dimension are displayed. Figure 5 provides an illustrative example of these types of figures in the case of sPLS. These graphical outputs proposed by González et al. (2008a) improve the interpretability of the results in the following manner. First they allow for the identification of correlated gene subsets from each data set, which are either up or down regulated. Second they help revealing the correlation between gene subsets from both data sets (by superimposing both graphics). And third they help relating these correlated subsets to the associated tumor cell lines by combining the information contained in Fig. 5 and Fig. 4 **(a)**. For example, we can make the assumption that the genes which were selected on the second sPLS dimension for both data sets should help discriminating melanoma tumors from the other cell lines.
In our case, these types of graphics usually show that there is few overlap between the gene selections in dimensions 1, 2 or 3. This means that each selection focus on a specific aspect of the data set (a specific tumor), and that the loading vectors are orthogonal ($cor(a_s, a_r) = 0$, $cor(b_s, b_r) = 0$, $r < s$). This valuable property is still kept in the sparse methods (sPLS, CCA-EN), which is not often the case (see the sparse PCA approaches, Zou and Hastie, 2005; Jolliffe et al., 2003; Shen and Huang, 2007). This results in a very small overlap between each gene list from each CCA-EN or sPLS dimension (Table 2). In fact, only 0 to 2 genes are overlapping the dimensions 1-2 and 1-3 in X, and between 1 to 13 genes in Y for both approaches. On the other hand, there is no orthogonality between CIA loadings vectors, leading to a high number of genes that are overlapping.

Comparisons of the gene lists.

Based on the interpretation of the cell line clusters (paragraph 3.3), our analysis of gene clusters relied on three sets of gene lists (3 methods $\times$ 2

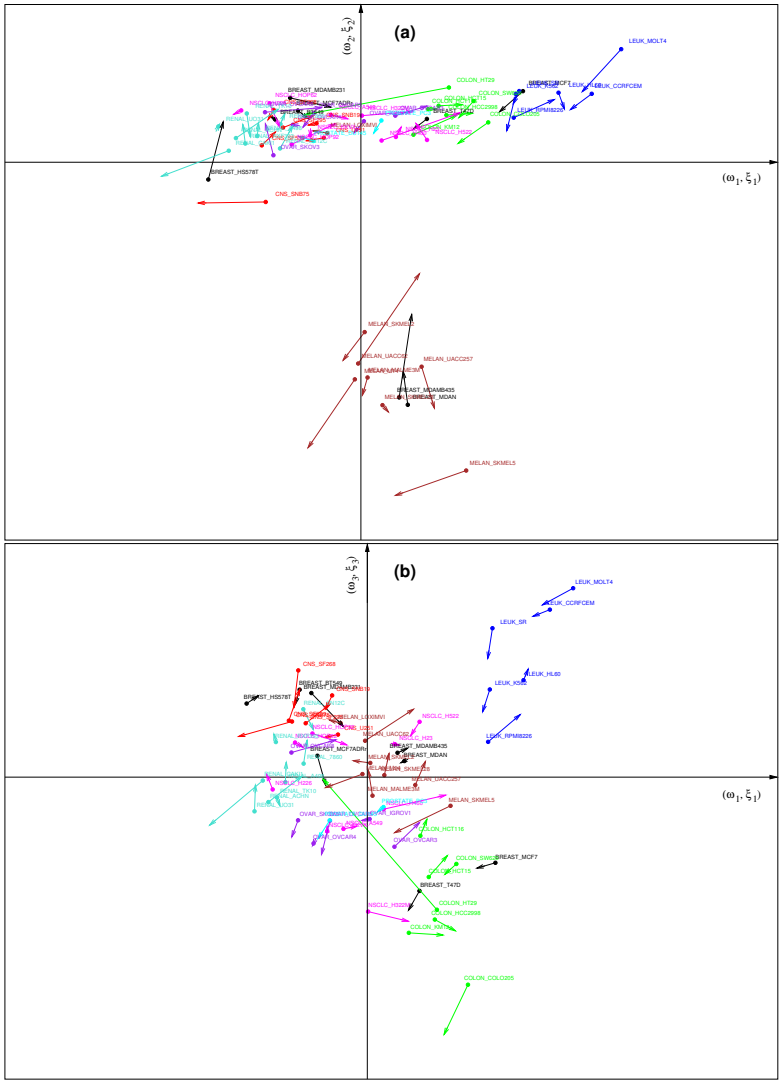


Figure 4: Graphical representations of the cell lines. sPLS latent variables are displayed in a superimposed manner, where the start of the arrow show the location of the Ross samples, and the tip the Staunton samples. The first and second axis (first and third) are shown in **(a)** and **(b)**.

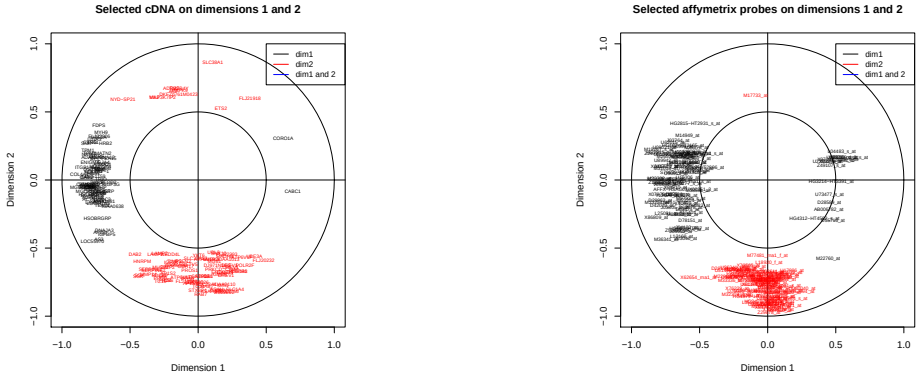


Figure 5: Graphical representation of the genes selected on the first two axes with sPLS. The coordinates of each gene are obtained by computing the correlation between (ξ_1, ξ_2) (resp. (ω_1, ω_2)) and the original Ross (resp. Staunton) data set. Selected cDNAs from the Ross data set (left) or selected Affymetrix probes from the Staunton data set (right) are displayed.

Table 2: Number of genes commonly selected between all dimensions for each approach.

	X=Ross-cDNA				Y=Staunton-Affymetrix			
	dim 1-2	dim 1-3	dim 2-3	dim 1-2-3	dim 1-2	dim 1-3	dim 2-3	dim 1-2-3
CCA-EN	0	2	2	0	1	3	13	1
CIA	20	17	31	2	14	21	24	1
sPLS	0	0	2	0	0	8	1	0

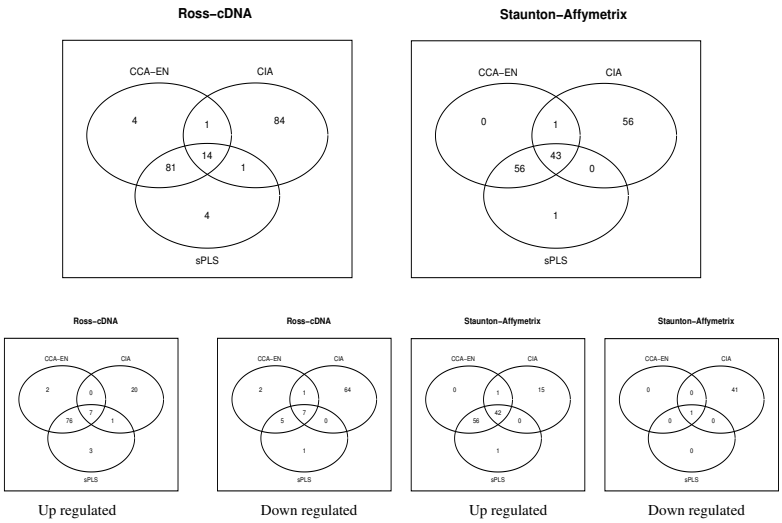


Figure 6: Venn diagrams for 100 selected genes associated to melanoma *vs.* the other cell lines for each data set (top) and by identifying up and down regulated genes in these lists (bottom).

data sets = 6 lists of 100 genes per set):

- **Set 1:** the lists associated with the separation of cell lines with epithelial (mainly LE and CO) versus mesenchymal (mainly RE and CNS) characteristics (CCA-EN axis 2, CIA and sPLS axes 1),
- **Set 2:** the lists associated with the separation of the melanoma cell lines (ME, BR_MDAN and BR_MDAMB435) from the other cell lines (CCA-EN axis 1, CIA and sPLS axes 2),
- **Set 3:** the lists associated with the separation of the LE cell lines from the CO cell lines (axis 3 of each method).

First, we evaluated for each set of gene lists and for each data set the number of genes that were selected in common by the different methods. Figure 6 displays the Venn diagrams for the lists of genes associated with melanoma *vs.* the other cell lines in dimension 1 for CCA-EN and in dimension 2 for CIA and sPLS.
For each set of gene lists, the Venn diagrams revealed a very strong similarity between the gene selections obtained by CCA-EN and sPLS, whereas CIA seemed to select other genes linked to the cell lines. Note that the same trend was observed if more than 100 variables were selected on each dimension.

Second, we evaluated for each dimension from each method the degree of overlap between the two data sets. In fact, it could be expected from the canonical methods that they identify correlations between measurements obtained from the two platforms when these correspond to the same gene. To evaluate this aspect, the identifiers of the features from each platform were mapped to unique gene identifiers using Ingenuity Pathway Analysis Application (IPA, see Supplemental Table). For all three dimensions, CCA-EN and sPLS selected approximately 20 features from the Ross and Staunton data sets that corresponded to identical genes. On the other hand, CIA selected 15 to 17 genes that were common to the two data sets.

We obtained heatmaps (Sup. Fig. 7, 8 and 9) for each of the 18 gene lists. For all heatmaps, we used the clustering of the individuals obtained with the Ross and Staunton data sets, presented in Fig.2. These heatmaps

illustrate well the general finding that CCA-EN and sPLS are most similar, and that CIA tends to select genes with a higher variance across all cell lines compared to CCA - EN and sPLS.
The 3 sets of gene lists were loaded into IPA along with their corresponding log ratios and we focused on 1) *biological functions* that were significantly over-represented (right-tailed Fisher's exact test) in the gene lists compared to the initial sets of genes (1,375 and 1,517 genes for the Ross and Staunton data sets respectively), 2) *canonical pathways* in which the genes from the lists were significantly over-represented compared to the genes in the initial sets and 3) the first *networks* generated by IPA from the gene lists. These networks are build by combining the genes into small (35 molecules maximum) networks that maximize their specific connectivity (Calvano et al., 2005) which result in highly-interconnected networks. The main results from these analyses are presented below for each set.

Set 1: Epithelial-Mesenchymal Transition (EMT). As previously described for a CIA analysis (Culhane et al., 2003), axes 1 (CIA and sPLS) or 2 (CCA-EN) of the 3 methods distinguished cell lines with epithelial characteristics (mainly CO and LE) from cell lines with stromal/mesenchymal characteristics (mainly RE and CNS). The epithelial to mesenchymal transition (EMT) is a key process underlying various tissue remodeling events during embryonic development. The EMT is thought to be also involved in establishing the metastatic potential of carcinoma cells (Yang and Weinberg, 2008). During the EMT, cells acquire morphological and biochemical characteristics that enables them to limit their contacts with neighboring cells and to invade the extracellular matrix. Studying the events underlying this process is thus of primary importance to better understand tumor malignancy.
The most significant biological functions identified in common by the three methods ($p < 0.001$ for each method) were:

- Cellular movement, skeletal and muscular system development and function, tissue development, cell-to-cell signaling and interaction, cellular assembly and organization and cancer for the Ross data set
- Cell morphology, cellular movement, cell death, cancer, reproductive

system disease, cell-to-cell signaling and interaction, connective tissue development and function, cellular function and maintenance, cardiovascular system development and function, renal and urological system development and function and cellular development for the Staunton data set

The lists of genes involved in these biological functions are available as Supplemental Table). First, this illustrates well the complementarity of the two data sets, which interrogate very different sets of genes (see Culhane et al., 2003 for such a comparison) and may thus identify complementary aspects of the same biological process. Second, most of the biological functions identified are highly relevant to the EMT transition which involves modifications of the connective tissue and of cell morphology, cell movement and cell-to-cell interactions in particular. Genes involved in skeletal and muscular system development were found to be more highly expressed in stromal/mesenchymal cell lines and is consistent with previous observations (Ross et al., 2000; Tso et al., 2006). Similarly, genes involved in the function "reproductive system disease" were mostly over expressed in stromal/mesenchymal cell lines and were mainly associated with breast cancer cell lines biological functions. This is consistent with the presence of most breast cancer cell lines on the stromal/mesenchymal side of the corresponding axes. Other biological functions were more specifically identified by CIA or CCA-EN/sPLS. Generally, the latter two methods identified the same biological functions, which is consistent with the similarity of their gene selections. However, CIA systematically identified (sometimes many) more highly significant biological functions compared to CCA-EN/sPLS (*e.g.* for the Ross data set, CCA-EN and sPLS identified 7 functions with $p < 0.001$ while CIA identified 21 functions using the same threshold). Since many of these functions were found significant for 2 to all 3 sets of gene lists, this likely reflects the redundancy in gene selections among the CIA axes. Thus, while some of these additional biological functions evidenced by CIA may be relevant, their interpretation may also be misled by less specific findings. This hypothesis was strengthened when we focused on the canonical pathways identified by IPA analysis. CCA-EN and sPLS both found that the integrin and the actin cytoskeleton pathways contained a significantly higher

number of genes that were over expressed in RE and CNS cell lines compared to LE and CO than could be expected by chance. This finding was consistent between the two data sets. These two central pathways in cell movement, which appear highly relevant to the EMT, displayed much higher p-values for the analysis of the gene lists selected by CIA. It is thus likely that less specific genes contained in the CIA gene selections limit the enrichment of a sufficient number of genes in a given pathway to yield low enough p-values.
Finally, the first networks identified by IPA for all three methods were highly connected and were associated with cellular movement for both data sets and in addition with cell-to-cell signaling and interaction for the Ross data set. Interestingly, all six networks pointed to the ERK (extracellular-signal-regulated kinase) signaling pathway as a central player in the gene expression modulations that were selected, which is consistent with its known role in cell migration (Juliano et al., 2004). However, the CIA network for the Ross data set failed to identify the integrin pathway as an upstream regulator of ERK. Merging the first 3 networks from the 3 canonical methods for each data set yielded two highly similar networks (Supplemental Figures 10 and 11). However, only the network built from the Satunton data set highlighted the transforming growth factor$-\beta$ (TGF$-\beta$) pathway which is thought to be a primary inducer of the EMT (Yang and Weinberg, 2008). Despite this difference, the most connected nodes (including integrins α and β alpha-actinin, connective tissue growth factor, fibronectin 1, SERPINE1, plasminogen activator urokinase, Ras or ERK) were found in both networks. These likely represent central players in establishing the different phenotypes of LE and CO cell lines on one hand and of RE and CNS cell lines on the other hand.

Set 2: Melanoma cell lines. Axis 1 of CCA and axes 2 of CIA and PLS clearly separate all except one (LOXIMVI) melanoma cell lines, along with the melanoma metastasis BR_MDAN and BR_MDAMB435 from all other cell lines. The melanoma cell line LOXIMVI has previously been shown to lack melanine and several typical markers of melanoma cells (Stinson et al., 1992), which likely explains its absence in the cluster of ME cell lines. For

these axes, the selections made by CCA-EN and sPLS are almost identical for the two data sets (only 1 and 5 genes specific to each method for the Staunton and the Ross data sets respectively).
For this cluster, less significant biological functions were identified compared to Set 1 and these differed substantially between CIA and the other two methods. The most significant biological functions ($p < 0.001$ for both methods) identified by CCA-EN/ sPLS were:

- Molecular transport, amino acid metabolism and small molecule biochemistry for the Ross data set
- Hair and skin development and function, amino acid metabolism, cellular development, small molecule biochemistry, cell morphology, dermatological diseases and conditions, nervous system development and function

On the other hand, CIA identified the following significant biological functions ($p < 0.001$):

- Cancer, reproductive system disease, cellular movement and cell morphology for the Ross data set
- Cellular growth and proliferation, cancer, hair and skin development and function, reproductive system disease, amino acid metabolism, cell morphology, cellular assembly and organization, ophthalmic disease and small molecule biochemistry for the Staunton data set

As for Set 1, CIA identified more biological functions than CCA-EN/sPLS but some, such as "cancer", appear less specific and are common to all three sets of gene lists. Overall, the biological functions identified by the three methods appear relevant to the characterization of melanoma cell lines, particularly those related to skin biology. The categories related to amino acid metabolism (including small molecule biochemistry and molecular transport which contains many genes involved in amino acid transport and metabolism) are likely found because ME cell lines are characterized by melanin synthesis which involves the amino acids tyrosine and cysteine. Similarly to Set 1, CCA-EN/ sPLS identified more significant canonical pathways compared to CIA which allowed a more precise understanding of

the gene lists selected by these two methods. In particular, CCA-EN/sPLS identified glycosphingolipid biosynthesis pathways from both the Ross (ganglioside biosynthesis only) and the Staunton data sets (ganglioside and globosid biosynthesis pathways). Melanoma tumors are known to be rich in these glycosphingolipids (Portoukalian et al., 1979). Indeed, their presence at the cell membrane makes them interesting targets for immunotherapy and vaccination strategies (Fredman et al., 2003). Noticeably, the tyrosine metabolism pathway was identified by all three methods ($p < 0.05$) in the Staunton data set but only by CCA-EN/sPLS in the Ross data set ($p < 0.05$). Genes involved in this pathway included tyrosinase, tyrosinase related proteins 1 and 2 and dopachrome tautomerase which are all involved in melanin biosynthesis and were found over expressed in melanoma cell lines accordingly.
Finally, the first networks generated by IPA from the CCA-EN/sPLS and the CIA gene lists pointed to differential activities or expression of several components of signaling pathways including TGF-β, PDGF, TNF, Mek, Erk, Mapk, Ras, PKA, PKCδ, Jnk, AP1, PI3K or Akt in melanoma cell lines compared to the other cell lines. These networks, especially those obtained from the Staunton data set, also highlighted several markers used for the diagnosis of melanomas including the over expressed MITF, Vimentin, S-100A1, S-100B and Melan-A and the under expressed keratins 7, 8, 18 and 19.

Set 3: Leukemia cell lines compared to colon tumor cell lines. The axes 3 from each of the three canonical methods separated the LE from the CO cell lines highlighting that these two groups could also be distinguished through gene expression profiles of selected genes from both data sets.
For the Ross data set, CIA found only one significant biological function (tissue development) that had not been found significant at the 0.001 threshold for Sets 2 and 3. Most of the genes in this category were expressed at lower levels in LE compared to CO cell lines and were implicated in the adhesion of epithelial cells or tissue and in the formation and assembly of extracellular matrix. CCA-EN and sPLS identified the hematological and immunological disease categories as relevant biological functions that sep-

arate the LE from the CO cell lines for the Ross data set. In addition, they identified the cell death category that was also found for the Staunton gene lists of Set 1 but the genes implicated in this biological function were almost completely different between Set 1 and Set 3. For the Staunton data set, CCA-EN alone identified a set of genes implicated in embryonic development that were over expressed in CO cell lines compared to LE cell lines (except CXCR4 that was over expressed in LE compared to CO). Interestingly, all three methods identified a set of three genes implicated in severe combined immunodeficiency (CD3D, IL2RG and ZAP70) that were up regulated in LE compared to CO cells.
Surprisingly, CIA seemed to identify many more canonical pathways for the Ross data set compared to CCA-EN and sPLS. Indeed these were all specific metabolic pathways involving the same three isoforms of poorly specific aldehyde dehydrogenase. sPLS alone identified the tight junction signaling pathway which included in particular Claudin 4 (CLDN4) and Zona occludens 1 (ZO1) that are strongly expressed in CO cell lines but not in LE cell lines and are key components of the tight junctions between epithelial cells. A similar bias in canonical pathway identification was observed for the Staunton data set for which CCA-EN and sPLS had selected two aldehyde dehydrogenases along with other enzymes involved in several metabolic pathways.
The first networks found by IPA for the Ross data set were mainly focused on genes involved in cell-to-cell signaling and interaction and in cellular movement, assembly and organization. In particular, most of these genes were components of the cytoskeleton, of the basement membrane or of cell-cell junctions. They were also involved in cell-cell contacts or in cell migration and adhesion. Most of them were expressed at much higher levels in CO versus LE cell lines, which is consistent with the typical epithelial characteristics of the colon tumor cell lines compared to the leukemia cell lines. For the Staunton data set, the first networks identified by IPA were also mainly focused on cell-to-cell signaling and interaction and on cellular movement. Overall, these results highlighted the fact that the CO cell lines are much more characteristic of an epithelium than the LE cell lines.

Conclusion

The analysis of the NCI60 data sets with CCA-EN, CIA and sPLS evidenced the main differences between these methods.

CIA. CIA does not propose a built-in variable selection procedure and requires a two-step analysis to perform variable selection. The main individual effects were identified. However, the loadings or weight vectors obtained were not orthogonal, in contrary to CCA-EN and sPLS. This resulted in some redundancy in the gene selections on the first three axes, which may be a limitation for the biological interpretation, as there may be less specific genes related to some cell lines types that were identified.
The gene selections obtained on each dimension generally led to interpretations that were overall similar to those obtained with CCA-EN and sPLS. However, the interpretations of the gene selections were clearly affected by genes selected on several axes, leading to less specific results.

CCA-EN. CCA-EN first captured the main robust effect on the individuals that is present in the two data sets. Consequently, it may hide strongest individual effects that are present in only one data set, but bring robust results.
We observed a strong similarity between CCA-EN and sPLS in the gene selections, except that the axes were permuted. In fact, we believe that CCA-EN can be considered as a sparse PLS variant with a canonical mode. Indeed, the elastic net is approximated with a univariate threshold, similar to a soft-thresholding penalization, and the whole algorithm uses PLS and not CCA computations, which explains why the canonical correlations do not monotonically decrease. The only difference that distinguishes sPLS canonical mode from CCA-EN is the initialization of the algorithm for each dimension. CCA-EN maximizes the correlation between the latent variables, whereas sPLS maximizes the covariance.

sPLS. We found that sPLS makes a good compromise between all these approaches. It includes variable selection and the loading vectors are or-

thogonal. Apart from the fact that sPLS and CCA-EN do not order the axis in the same manner, both approaches were highly comparable, except for slight but significant differences when studying LE *vs.* CO (axes 3). In this particular case, the resulting gene lists clearly provided complementary information.

We believe that all approaches are easy to use and fast to compute. These approaches would benefit from the development of an `R` package that could harmonize their inputs and outputs to facilitate their use and their comparison. Based on the present study, we would primarily recommend the use of CCA-EN or sPLS when gene selection is an issue. Like CCA-EN, sPLS includes a built-in variable selection procedure but captured subtle individual effects. Therefore, the choice of one of these methods would take into consideration the fundamental difference between them in the building of the first axes.

References

Butte, A., Tamayo, P., Slonim, D., Golub, T., and Kohane, I. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proceedings of the National Academy of Sciences, page 220392197.

Bylesjö, M., Eriksson, D., Kusano, M., Moritz, T., and Trygg, J. (2007). Data integration in plant biology: the o2pls method for combined modeling of transcript and metabolite data. The Plant Journal, 52:1181–1191.

Calvano, S., Xiao, W., Richards, D., Felciano, R., Baker, H., Cho, R., Chen, R., Brownstein, B., Cobb, J., Tschoeke, S., et al. (2005). A network-based analysis of systemic inflammation in humans. NATURE-LONDON-, 437(7061):1032.

Combes, S., González, I., Déjean, S., Baccini, A., Jehl, N., Juin, H., Cauquil, L., and Batrice Gabinaud, Franois Lebas, C. L. (2008). Relationships between sensorial and physicochemical measurements in meat of rabbit from three different breeding systems using canonical correlation analysis. Meat Science, in press.

Culhane, A., Perriere, G., and Higgins, D. (2003). Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. BMC Bioinformatics, 4(1):59.

de Jong, S. (1993). Simpls: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 18:251–263.

Doledec, S. and Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biology, 31(3):277–294.

Fredman, P., Hedberg, K., and Brezicka, T. (2003). Gangliosides as Therapeutic Targets for Cancer. BioDrugs, 17(3):155.

Gittins, R. (1985). Canonical Analysis: A Review with Applications in Ecology. Springer-Verlag.

González, I., Déjean, S., Martin, P., Goncalves, O., Besse, P., and Baccini, A. (2008a). Highlighting Relationships Between Heterogeneous Biological Data Through Graphical Displays Based On Regularized Canonical Correlation Analysis. Technical report, Université de Toulouse.

González, I., Déjean, S., Martin, P. G. P., and Baccini, A. (2008b). Cca: An r package to extend canonical correlation analysis. Journal of Statistical Software, 23(12).

Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 28:321–377.

Jolliffe, I., Trendafilov, N., and Uddin, M. (2003). A Modified Principal Component Technique Based on the LASSO. Journal of Computational & Graphical Statistics, 12(3):531–547.

Juliano, R., Reddig, P., Alahari, S., Edin, M., Howe, A., and Aplin, A. (2004). Integrin regulation of cell signalling and motility. Biochem Soc Trans, 32:443–446.

Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., and Besse, P. (2008). Sparse PLS: Variable Selection when Integrating Omics data. Technical report, Université de Toulouse et Institut National de la Recherche Agronomique.

Lorber, A., Wangen, L., and Kowalski, B. (1987). A theoretical foundation for the PLS algorithm. Journal of Chemometrics, 1(19-31):13.

Portoukalian, J., Zwingelstein, G., and Dore, J. (1979). Lipid composition of human malignant melanoma tumors at various levels of malignant growth. FEBS Journal, 94(1):19.

Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the RV-coefficient. Applied Statistics, 25(3):257–265.

Ross, D., Scherf, U., Eisen, M., Perou, C., Rees, C., Spellman, P., Iyer, V., Jeffrey, S., Van de Rijn, M., Waltham, M., et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet, 24(3):227–35.

Scherf, U., Ross, D., Waltham, M., Smith, L., Lee, J., Tanabe, L., Kohn, K., Reinhold, W., Myers, T., Andrews, D., et al. (2000). A gene expression database for the molecular pharmacology of cancer. Nat Genet, 24(3):236–244.

Shen, H. and Huang, J. Z. (2007). Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis, to appear.

Staunton, J., Slonim, D., Coller, H., Tamayo, P., Angelo, M., Park, J., Scherf, U., Lee, J., Reinhold, W., Weinstein, J., et al. (2001). Chemosensitivity prediction by transcriptional profiling. Proceedings of the National Academy of Sciences, 98(19):10787.

Stinson, S., Alley, M., Kopp, W., Fiebig, H., Mullendore, L., Pittman, A., Kenney, S., Keller, J., and Boyd, M. (1992). Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen. Anticancer Res, 12(4):1035–53.

Tenenhaus, M. (1998). La régression PLS: théorie et pratique. Editions Technip.

Thioulouse, J., Chessel, D., Dole´ dec, S., and Olivier, J. (1997). ADE-4: a multivariate analysis and graphical display software. Statistics and Computing, 7(1):75–83.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 58(1):267–288.

Tso, C., Shintaku, P., Chen, J., Liu, Q., Liu, J., Chen, Z., Yoshimoto, K., Mischel, P., Cloughesy, T., Liau, L., et al. (2006). Primary Glioblastomas Express Mesenchymal Stem-Like Properties. Molecular Cancer Research, 4(9):607.

Vijayendran, C., Barsch, A., Friehs, K., Niehaus, K., Becker, A., and Flaschel, E. (2008). Perceiving molecular evolution processes in Escherichia coli by comprehensive metabolite and gene expression profiling. Genome Biology, 9(4):R72.

Vinod, H. D. (1976). Canonical ridge and econometrics of joint production. Journal of Econometrics, 4(2):147–166.

Waaijenborg, S., de Witt Hamer, V., Philip, C., and Zwinderman, A. (2008). Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis. Statistical Applications in Genetics and Molecular Biology, 7(1):3.

Wegelin, J. (2000). A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Technical Report 371, Department of Statistics, University of Washington, Seattle.

Wold, H. (1966). Multivariate Analysis. Academic Press, New York, Wiley, krishnaiah, p.r. (ed.) edition.

Wold, S., Eriksson, L., Trygg, J., and Kettaneh, N. (2004). The PLS method–partial least squares projections to latent structures–and its applications in industrial RDP (research, development, and production). Technical report, Umea University.

Yang, J. and Weinberg, R. (2008). Epithelial-Mesenchymal Transition: At the Crossroads of Development and Tumor Metastasis. Developmental Cell, 14(6):818–829.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B, 67(2):301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. Journal of Computational and Graphical Statistics, 15(2):265–286.

Appendix

A PLS algorithm (canonical mode)

1. $X_0 = X$, $Y_0 = Y$
2. For $h = 1 \dots H$:
   (a) Initialize
   ξ_h = first column of X_{h-1} $\quad \omega_h$ = first column of Y_{h-1}
   (b) Until convergence of a_h:
      i. $a_h = X'_{h-1}\xi_h / \xi'_h\xi_h$, norm a_h
      ii. $\xi_h = X_{h-1}a_h$, norm ξ_h
      iii. $b_h = Y'_{h-1}\xi_h / \xi'_h\xi_h$, norm b_h
      iv. $\omega_h = Y_{h-1}b_h$, norm ω_h
   (c) $c_h = X'_{h-1}\xi_h \qquad e_h = Y'_{h-1}\omega_h$
   (d) $X_h = X_{h-1} - \xi_h c'_h \quad Y_h = Y_{h-1} - \omega_h e'_h$

Step (c) computes the regression coefficients of the matrices X_{h-1} and Y_{h-1} on the latent variables ξ_h and ω_h.
Step (d) computes the deflated (residual) matrices.

B sparse PLS algorithm (canonical mode)

Sparse PLS initializes step (a) in PLS by extracting the first pair of singular vectors (a_h, b_h) of the crossproduct $X'_{h-1}Y_{h-1}$, which includes variation in both X and Y and the correlation between the two sets.
The two loading vectors (a_h, b_h) are then computed with penalizations λ_1

and λ_2 in step (b), and the latent vectors (ξ_h, ω_h) are then computed, where $g(y) = sign(y)(|y| - \lambda)_+$ is the soft-thresholding penalty function.

1. $X_0 = X \qquad Y_0 = Y$
2. For $h = 1 \dots H$:
   (a) Set $\tilde{M}_{h-1} = X'_{h-1} Y_{h-1}$, decompose $\tilde{M}_{h-1}$ and extract the first pair of singular vectors $a_{old} = a_h$ and $b_{old} = b_h$
   (b) Until convergence of a_{new} and b_{new}:
      i. $a_{new} = g_{\lambda_1}(\tilde{M}_{h-1} b_{old})$, norm a_{new}
      ii. $b_{new} = g_{\lambda_2}(\tilde{M}'_{h-1} a_{old})$, norm b_{new}
      iii. $a_{old} = a_{new}$, $b_{old} = b_{new}$
   (c) $\xi_h = X_{h-1} a_{new}$
   $\omega_h = Y_{h-1} b_{new}$
   (d) $c_h = X'_{h-1} \xi_h \quad e_h = Y'_{h-1} \omega_h$
   (e) $X_h = X_{h-1} - \xi_h c'_h \quad Y_h = Y_{h-1} - \omega_h e'_h$

C Canonical Correlation Analysis with Elastic Net penalization

CCA-EN initializes step (a) in PLS by setting $\xi_h = X^i_{h-1}$ and $\omega_h = Y^j_{h-1}$ such that $cor(X^i_{h-1}, Y^j_{h-1})$ is maximised, for $i = 1 \dots p$ and $j = 1 \dots q$. Hence this algorithm aims at maximising the correlation (rather than the covariance for PLS and sPLS).
The approximation on Elastic Net penalization finally consists in introducing lasso penalizations, as in sparse PLS, which makes both algorithms similar.

1. $X_0 = X \qquad Y_0 = Y$
2. For $h = 1 \dots H$:
   (a) Set $\xi_h = X^i_{h-1}$ and $\omega_h = Y^j_{h-1}$ such that $cor(X^i_{h-1}, Y^j_{h-1})$ is maximised
   $a_{new} = X'_{h-1} \xi_h / \xi'_h \xi_h \quad b_{new} = Y'_{h-1} \xi_h / \xi'_h \xi_h$, norm a_{new} and b_{new}

(b) Until convergence of a_{new} and b_{new}:

i. $a_{new} = g_{\lambda_1}(Y_{h-1}b_{old})$, norm a_{new}

ii. $b_{new} = g_{\lambda_2}(X_{h-1}a_{old})$, norm b_{new}

iii. $a_{old} = a_{new}$, $b_{old} = b_{new}$

(c) $\xi_h = X_{h-1}a_{new}$, norm ξ_h
$\omega_h = Y_{h-1}b_{new}$ norm ω_h

(d) $c_h = X'_{h-1}\xi_h \quad e_h = Y'_{h-1}\omega_h$

(e) $X_h = X_{h-1} - \xi_h c'_h \quad Y_h = Y_{h-1} - \omega_h e'_h$

D Supplemental figures

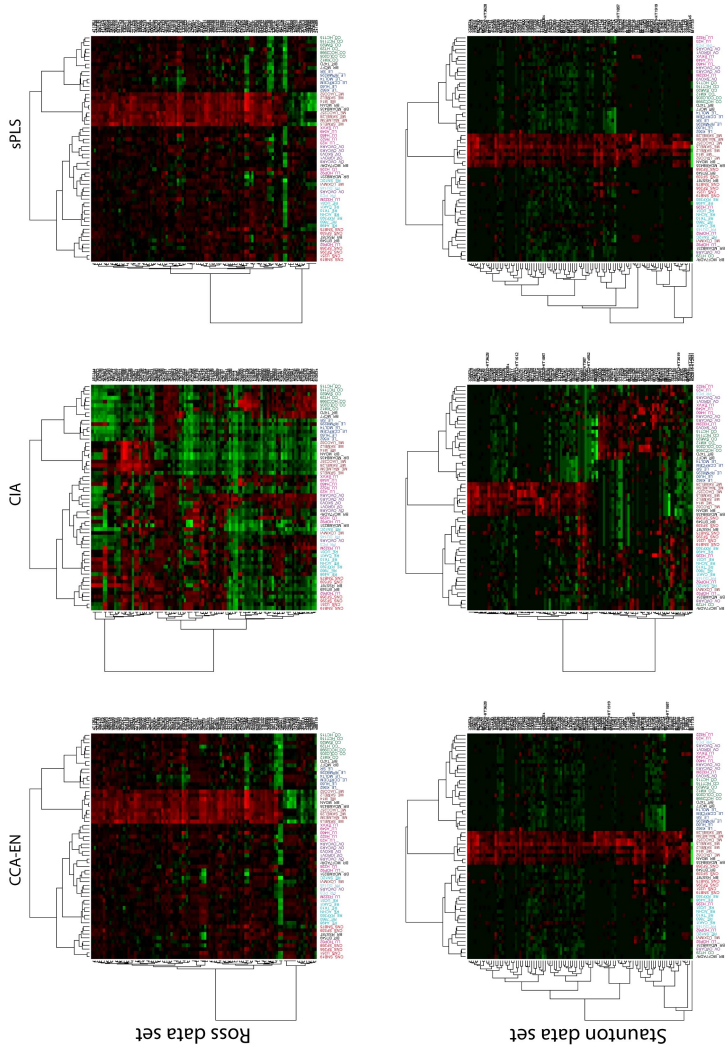


Figure 7: Heat map displays of hierarchical clustering results with the Ward method and correlation distance with genes in lines and cell lines in columns. The red (green) colour represents over-expressed (under-expressed) genes. Selection of 100 genes on dimension 2 (resp. 1) for CIA and sPLS (resp. CCA-EN) that separate melanoma *vs.* the other cell lines.

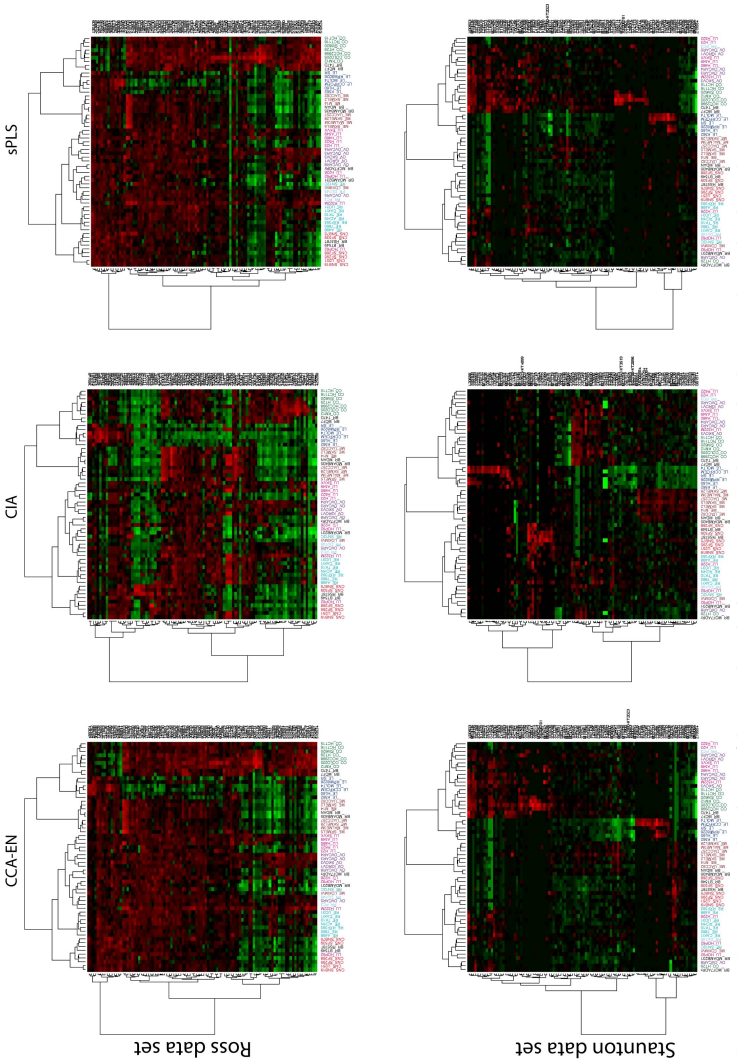


Figure 8: Heat map displays of hierarchical clustering results with the Ward method and correlation distance with genes in lines and cell lines in columns. The red (green) colour represents over-expressed (under-expressed) genes. Selection of 100 genes on dimension 3 (resp. 2) for CIA and sPLS (resp. CCA-EN) that separate LE *vs.* CO.

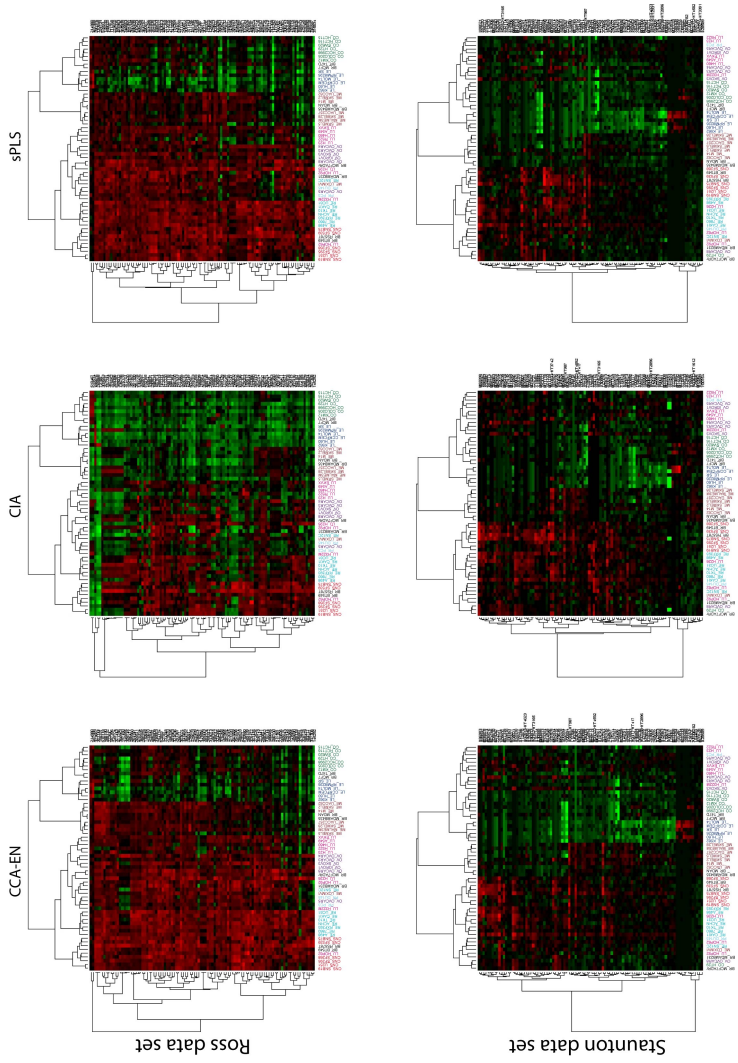


Figure 9: Heat map displays of hierarchical clustering results with the Ward method and correlation distance with genes in lines and cell lines in columns. The red (green) colour represents over-expressed (under-expressed) genes. Selection of 100 genes on dimension 1 (resp. 2) for CIA and sPLS (resp. CCA-EN) that separate epithelial-like *vs.* mesenchymal-like.

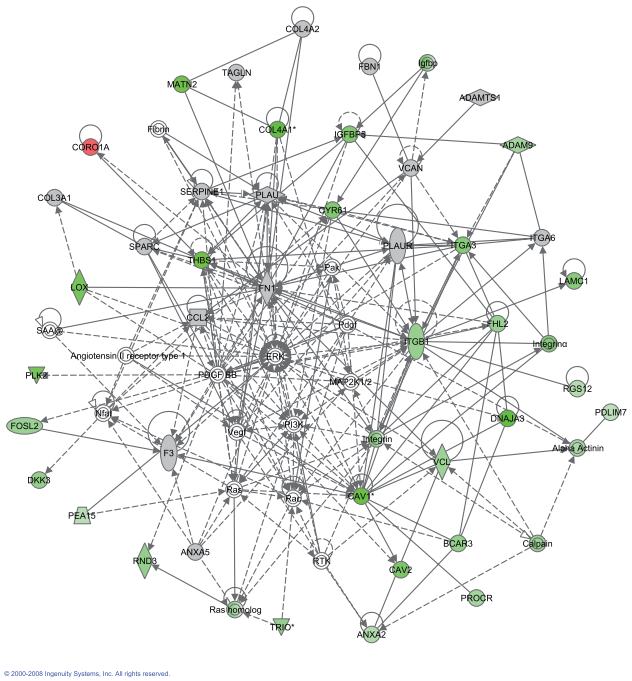


Figure 10: Molecular network obtained from the Ross gene lists from Set 1. For each canonical method (CCA-EN, CIA or sPLS), molecular networks were built from the Ross gene lists (focus genes) of Set 1 using Ingenuity Pathways Analysis (IPA, www.ingenuity.com). The first networks obtained from each method were merged into the presented network. Green and red colors indicate under- and over-expressions respectively in the LE/CO cell lines compared to the RE/CNS cell lines. Only the genes selected by sPLS have been colored in red or green. Genes colored in grey have been selected by CCA-EN or sPLS and all correspond to genes that are under-expressed in the LE/CO cell lines compared to the RE/CNS cell lines. Genes in white have been added by IPA based on their high connectivity with focus genes.

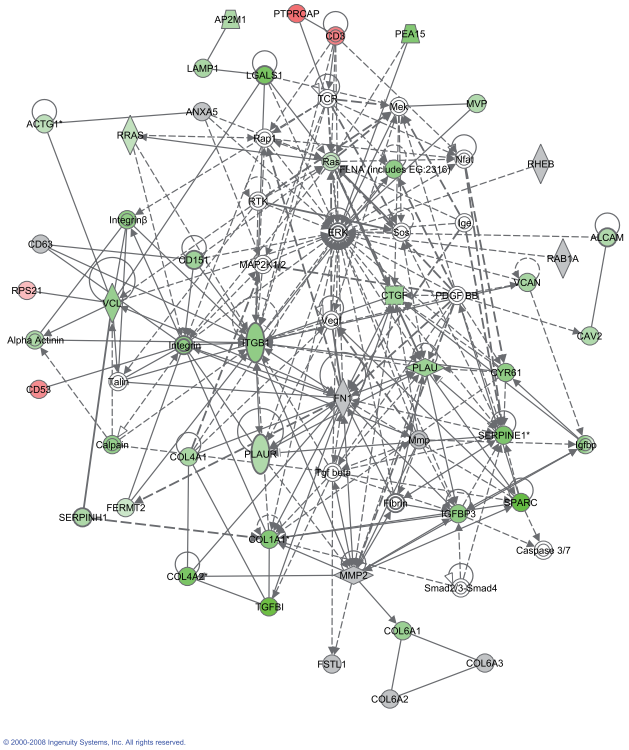


Figure 11: Molecular network obtained from the Staunton gene lists from Set 1. For each canonical method (CCA-EN, CIA or sPLS), molecular networks were built from the Staunton gene lists (focus genes) of Set 1 using Ingenuity Pathways Analysis (IPA). The first networks obtained from each method were merged into the presented network. Green and red colors indicate under- and over-expressions respectively in the LE/CO cell lines compared to the RE/CNS cell lines. Only the genes selected by sPLS have been colored in red or green. Genes colored in grey have been selected by CCA-EN or sPLS and all correspond to genes that are under-expressed in the LE/CO cell lines compared to the RE/CNS cell lines. Genes in white have been added by IPA based on their high connectivity with focus genes.

10. Bilan et perspectives

Nous avons montré dans cette partie que l'approche proposée sparse PLS répond à la question du biologiste dans un contexte d'intégration de variables omiques, et ceci à la fois dans un cadre de régression et dans un cadre d'analyse canonique. Ici aussi nous ne répondons pas à la question de la taille optimale des sélections de variables de chaque tableau, et laissons le biologiste libre de choisir en fonction des contraintes du jeu de données étudié.

Les programmes ont été développés de façon à en faciliter l'usage et l'utilisateur peut directement fixer le nombre de variables à sélectionner, plutôt que de paramétrer les constantes de pénalisation (comme proposé par Zou & Hastie (2005), dans le package `R elastic net`). L'ensemble de ces programmes feront partie d'un package `R` et `Bioconductor`, de manière à rendre cette approche la plus accessible possible au biologiste. Pour résoudre certains problèmes d'allocation de mémoire en `R`, il sera probablement nécessaire d'écrire une partie des programmes en Fortran ou C/C++.

D'un point de vue combinatoire, la sparse PLS ne semble pas être limitée par le nombre de variables de chaque tableau. Adaptée pour des données binaires de type 0-1, cette méthode pourrait être très utile dans le cadre de la modélisation de la sélection génomique en génétique animale. L'objectif ici est de mettre en relation des données phénotypiques, mesurées sur des centaines d'individus, avec des données de plusieurs milliers de SNP (*Single Nucleotide Polymorphism*), observés sur les mêmes individus. Dans ce cadre, on cherche à prédire un phénotype à partir de la meilleure sélection d'un sous-ensemble de SNP.

Quelques développements mathématiques seront nécessaires pour vraiment comprendre les différences observées entre la PLS-mode canonique et la CCA qui, lorsque les données sont centrées et réduites, devraient répondre à des objectifs similaires (dans un cas de petite dimension où $p + q < n$). La question ne semble pas avoir été abordée pour le moment dans la littérature statistique.

Dans le cadre des données omiques, il serait très utile pour les biologistes de développer un outil capable d'intégrer et sélectionner des variables issues de k jeux de données ($k > 2$). Dans ce domaine, l'analyse de tableaux multiples a connu beaucoup de développements et il existe une littérature très riche. Des approches comme kernel CCA (Scholkopf & Smola, 2001; Yamanishi *et al.*, 2003) ou l'analyse de co-inertie multiple (Chessel & Hanafi, 1996) pourraient notamment répondre à la question.

Enfin, il serait intéressant d'approfondir le *Dantzig selector* proposé par Candes & Tao (2007) qui, dans le cadre de la régression, estime $\hat{\beta}$ comme la solution du problème :

$$\min ||\beta||_1 \quad \text{sous la contrainte} \quad ||X'(y - X\beta)||_\infty \leq \lambda, \tag{10.1}$$

où λ est un paramètre à régler et $||X'r||_\infty = \sup_{0 \leq j \leq p} |X'r|$, avec $r = y - X\hat{\beta}$ le vecteur des résidus. La pénalisation l_1 permet de faire de la sélection de variables, tandis que la norme l_∞ contraint les résidus à prendre une petite valeur (proche du bruit), de façon à inclure dans le modèle des variables hautement correlées à y. Ce problème est convexe, mais nécessite cependant des calculs intensifs. James *et al.* (2007) ont récemment proposé de résoudre (10.1) par un algorithme similaire à LARS. La méthode reste cependant limitée lorsque p est très grand.

Troisième partie

Contribution à des développements en biologie.

11. Etude de la folliculogénèse chez le porc

Cet article se place dans le contexte d'une étude du développement folliculaire chez la truie, juste avant la phase d'ovulation. Nous étudions ici trois tailles de follicules : petits, moyens ou gros follicules. Le jeux de données est caractérisé par une très grande similarité biologique entre les petits et moyens follicules, avec les petits follicules représentant la classe minoritaire, tandis que les gros follicules se différencient facilement et constituent la classe majoritaire. Cette étude, commencée lors du stage de D.E.A (Lê Cao, 2005) et poursuivie en début de thèse, a tout de suite posé le problème du multiclasse déséquilibré. Nous avons aussi identifié la limite des méthodes filtres comme le test de Fisher, qui n'identifie comme gènes différentiellement exprimés que des gènes sur-exprimés pour les gros follicules, et donc peu informatifs concernant les autres classes.
Une méthode de type *wrapper* est appliquée (Balanced Random Forests de Chen *et al.* 2004) en plus d'un simple test de Fisher pour répondre à la question des biologistes. La validation de l'expérience biologique a été faite sur certains des gènes sélectionnés par PCR quantitative, tandis que les listes de gènes sélectionnés ont été étudiées grâce au logiciel Ingenuity Pathways Analysis pour identifier les réseaux, fonctions et voies métaboliques associés.

Cet article est publié dans la revue Reproduction (accepté le 28 avril 2008, sous presse).

REPRODUCTION
RESEARCH

In vivo gene expression in granulosa cells during pig terminal follicular development

A Bonnet, K A Lê Cao[1,3], M SanCristobal, F Benne, C Robert-Granié[1], G Law-So, S Fabre[2], P Besse[3], E De Billy, H Quesnel[4], F Hatey and G Tosser-Klopp

INRA, UMR 444, Génétique Cellulaire, F-31326 Castanet-Tolosan Cedex, France, [1]INRA, UR, Station d'Amélioration Génétique des Animaux, F-31326 Castanet-Tolosan Cedex, France, [2]INRA, UMR6175, Physiologie de la Reproduction et des Comportements, CNRS, Université de Tours, Haras Nationaux, F-37380 Nouzilly, France, [3]Université Paul Sabatier, UMR 5219, Institut de Mathématiques, F-31062 Toulouse Cedex 9, France and [4]INRA, UMR INRA/AgroCampus Rennes: Systèmes d'Elevage, Nutrition Animale et Humaine, F-35590 Saint Gilles, France

Correspondence should be addressed to A Bonnet; Email: agnes.bonnet@toulouse.inra.fr

Abstract

Ovarian antral follicular development is clearly dependent on pituitary gonadotrophins FSH and LH. Although the endocrine mechanism that controls ovarian folliculogenesis leading to ovulation is quite well understood, the detailed mechanisms and molecular determinants in the different follicular compartments remain to be clarified. The aim of this study was to identify the genes differentially expressed in pig granulosa cells along the terminal ovarian follicle growth, to gain a comprehensive view of these molecular mechanisms. First, we developed a specific micro-array using cDNAs from suppression subtractive hybridization libraries (345 contigs) obtained by comparison of three follicle size classes: small, medium and large antral healthy follicles. In a second step, a transcriptomic analysis using cDNA probes from these three follicle classes identified 79 differentially expressed transcripts along the terminal follicular growth and 26 predictive genes of size classes. The differential expression of 18 genes has been controlled using real-time PCR experiments validating the micro-array analysis. Finally, the integration of the data using Ingenuity Pathways Analysis identified five gene networks providing descriptive elements of the terminal follicular development. Specifically, we observed: (1) the down-expression of ribosomal protein genes, (2) the genes involved in lipid metabolism and (3) the down-expression of cell morphology and ion-binding genes. In conclusion, this study gives new insight into the gene expression during pig terminal follicular growth *in vivo* and suggested, in particular, a morphological change in pig granulosa cells accompanying terminal follicular growth.

Reproduction (2008) **136** 1–14

Introduction

The growth and development of ovarian follicles leading to ovulation require a series of coordinated events that lead to follicular somatic cell differentiation and oocyte development. It includes two successive periods, preantral (primordial, primary and secondary follicles) and antral (early antral, antral and preovulatory follicles) follicular developments. Follicles undergo morphological and functional changes as they progress towards ovulation. Among all growing follicles, only a small proportion of them (less than 1%) will ovulate. Most of them undergo a degenerative process called atresia that occurs at the different developmental stages. The antral follicular development depends on a complex regulatory network with endocrine regulation by pituitary gonadotrophins follicle-stimulating hormone (FSH) and luteinizing hormone (LH) but also with autocrine and paracrine pathways (Hsueh 1986) including the action of steroids and peptides (Hillier & Miro 1993, Drummond 2006). Those factors control follicular growth either directly or indirectly: for example, bone morphogenetic protein or the insulin-like growth factor (IGF) systems (Monget *et al.* 2002, Mazerbourg *et al.* 2003, Shimasaki *et al.* 2004) modify the sensitivity of follicular cells to FSH, and epidermal growth factor signalling network potentializes LH action (Hattori *et al.* 1995). Follicular development is thus a complex process and requires the coordinate expression of a large number of genes. The mechanisms underlying this development have been intensively studied mainly in granulosa cells because they constitute an important compartment in the mammalian ovarian follicle. They actively participate in the endocrine function of the ovaries by secreting oestradiol or progesterone under FSH or LH stimulation (Duda 1997).

In pigs, *in vitro* experiments or *in situ* hybridization have allowed the individual description of gene expression in granulosa cells during the growth of antral follicles. They referred mainly to the role of FSH, the IGF

ISSN 1470–1626 (paper) 1741–7899 (online)
DOI: 10.1530/REP-07-0312
Online version via www.reproduction-online.org

system and growth factors on gene expression regulation, and the expression of genes involved in steroidogenesis such as *CYP19A* (Chan & Tan 1987) and *STAR* (Balasubramanian *et al.* 1997). Despite continuing progress in the area of ovarian biology, many of the specific mechanisms involved in follicular development, including the initiation of primordial follicle growth, antrum formation, follicular growth/selection and follicular atresia, remain to be elucidated in greater detail. Then, global approaches have been undertaken to identify new genes involved in antral follicle maturation. Our previous data obtained on porcine granulosa cells *in vitro* by suppression subtractive hybridization (SSH) or differential display PCR suggested a role of FSH in extracellular matrix synthesis, chromatin remodelling, regulation of transcription activity and protection against atresia (Clouscard-Martinato *et al.* 1998, Bonnet *et al.* 2006*b*). Moreover, different cDNA libraries obtained from whole follicles of different size classes were generated to create a catalogue of differentially expressed genes along antral follicle development using sequence frequency in each library (Jiang *et al.* 2004). However, despite the development of DNA micro-array techniques in the pig (Tuggle *et al.* 2007), only few transcriptomic data are available concerning the ovarian function.

Thus, using the pig as a model and a transcriptomic approach, the aim of this study was to identify differentially expressed genes along terminal ovarian follicular development, before the onset of preovulatory LH surge. The experiment has focused especially on the time of expression of functional LH receptors in granulosa cells that occur in 4–5 mm follicle size and classically considered as a maturation marker in follicle development (May & Schomberg 1984). Then, we developed first a micro-array using cDNAs from SSH libraries obtained from granulosa cells isolated from three follicle size classes, small (1–2 mm), medium (3–4 mm) and large antral ($\geq$5 mm). In a second step, a transcriptomic analysis using cDNA probes from these three follicle classes identified transcripts differentially expressed along the terminal follicular growth and some of the gene networks associated with this process.

Results

Dedicated porcine micro-array tool construction

Our first goal was to construct a reliable tool to identify genes differentially expressed during terminal follicular development in porcine ovary by a global approach. Four SSH libraries were constructed (cf. Materials and Methods section) and screened leading to the selection of 1697 clones with the best 'differential' potential between small (SF), medium (MF) and large antral follicles (LF). These clones were sequenced, resulting in 1378 good quality sequences from which 441 sequences (35%) corresponded to the whole insert (presence of vector flanking sequences). The average insert sequence length was 640 bp.

All sequences were deposited in EMBL public database (accession numbers: CR939025–CR940296; CT971504–CT971571; CT990474–CT990531) and submitted to an assembly process (Sigenae contig assembly; Pig V3 p.sc.3, 30/01/2006). Our 1378 sequences were part of 345 contigs, 63 of which were new sequences. The contig analysis revealed a global redundancy of 75%. This redundancy was explained by the high proportion of sequences coming from only three genes, *CYP19A*, *GSTA* and *PGFS1* representing 15, 13.5 and 10% of the sequences respectively. There was no contig overlap between the forward and reverse libraries for SF versus LF comparison and around 10–18% for SF versus MF comparison.

The PCR products of the 1697 clones (345 genes/contigs) resulting from SSH experiments were spotted onto nylon membrane along with 1056 already sequenced PCR products (954 genes/contigs) to generate our micro-array platform (GEO accession number GPL3978).

Micro-array analysis

In order to identify genes whose expression differs along the terminal growth and maturation of porcine ovarian follicles, 14 complex cDNA radiolabelled probes synthesized from the three different size class follicles were hybridized onto our specific cDNA micro-array (1275 genes/contigs). Data (GEO accession number GSE5798 dataset) were pre-processed resulting in a list of 1564 expressed cDNAs, corresponding to 515 different genes or contigs of sequences assembled by SIGENAE.

A mixed linear model was applied to the 1564 cDNAs to quantify the hybridization signal intensity measured for each clone and for each complex probe in the function of follicle size class. The mean expression level for all genes was not statistically different between the follicle classes ($P=0.17$). By contrast, some genes had a significantly different expression level than others ($P<0.001$) and the significance of the gene–follicle class interaction ($P<0.001$) indicated some gene differential expression according to the follicle size classes. The biological variation represented 12% of the total variation while the variation between complex probe replicates and experimental variation corresponded respectively to 2 and 3% of the global variation.

The selection of significant differentially expressed genes was made through a *F*-test followed by a false discovery rate (FDR) adjustment on the *P* values (F analysis). The *P* values obtained after the FDR adjustment were very similar to the raw *P* values of the *F*-test, indicating a very low proportion of false positive (643 adjusted *P* value with FDR versus 705 raw F-statistics *P* values <0.002.). At the 0.2% level of significance, F analysis selected 643 cDNAs

corresponding to 75 known genes and 4 unknown contigs (Tables 1 and 2). The unsupervised hierarchical clustering shows that these 643 cDNAs separated the three follicle size classes in only two groups: LF versus MF and SF (Fig. 1). Among the 79 genes/contigs, 25 were overexpressed and 54 were down-expressed in LF group compared with SF/MF. However, when the expression of each selected gene was studied individually, five different expression profiles were observed (Tables 1 and 2), illustrated by the expression of glutathione S-transferase-α (*GSTA1*), *CYP19A*, *TUBA1B*, *EEF1A* and stathmin 1 (*STMN1*) (Fig. 2). Interestingly, numerous genes whose expression increased during the terminal follicular growth were implicated in glutathione metabolism (*GSTA1*, *GSTA2* and *MGST1*) and lipid metabolism (*CYP19A*, *AKR1C3*, *AKR1C4*, *HADHB*, *BDH2*, *CYB5*, *NR5A2* and *RTF1*). By contrast, the terminal follicular growth was notably accompanied with decreased expression of genes implicated in protein translation (16 subunits of ribosomal proteins, *EEF1A*), ion binding (calumenin (*CALU*), *SLC40A1*, calmodulin (*CALM1*) and *S100A11*) and cell shape (*TUBA1B*, *TUB5B*, *TUB7*, *VIM*, *CAPNS1*, *COF1*, smoothelin (*SMTN*), *STMN1*, *RPSA* and *DAG1*).

In order to identify a set of predictive genes that could help classify the follicles in their respective class, we have performed the random forest (RF) algorithm. The internal estimation of the generalization error was between 4.76 and 7.14% depending on the forests. A selection of the 120 most important and stable cDNAs with the Mean Decrease Gini importance measure gave an unsupervised hierarchical clustering allowing the separation of the three follicle size classes (Fig. 3). These 120 cDNAs corresponded to 24 known genes and two contigs and included 20 genes already selected by F analysis at the 0.2% level of significance (Tables 1 and 2). Among the six genes selected only by RF analysis, five were found below the 5% level of significance by F analysis. Only the differential expression of the *SFT2D2* gene was considered as non-significant ($P > 0.05$) by F analysis.

Differential expression validation by quantitative real-time PCR and in situ *hybridization*

In order to validate the micro-array analysis by quantitative real-time PCR, 18 differentially expressed genes selected by F and/or RF analysis were analysed (Table 3). Apart the *MT-CO1* gene, the differential

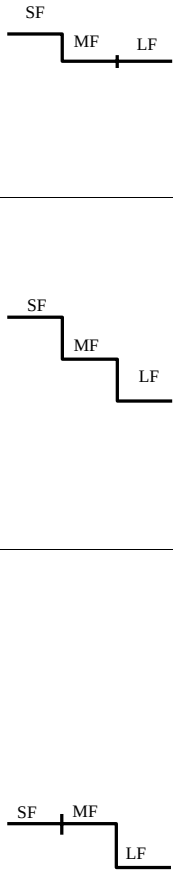
Table 1 Summary of significant overexpressed genes during pig follicular development.

HUGO name symbol	Gene description	FDR-adjusted *P* value	RF importance ($\times 10^{-3}$)	Maximum fold change	Expression profile
HADHB	Hydroxyacyl-coenzyme A dehydrogenase/3-ketoacyl-coenzyme A thiolase/enoyl-coenzyme A hydratase, β-subunit	‡	49.5	3.34	
PSMC2	Proteasome (prosome, macropain) 26S subunit, ATPase, 2	‡	47.4	2.73	
GSTA1*	Glutathione S-transferase A1	‡	222	2.58	
CTSL	Cathepsin L	‡		2.38	
HSPA8*	Heat shock 70 kDa protein 8	‡		2.38	
MGST1*	Microsomal glutathione S-transferase 1	‡		2.19	
GSTA2	Glutathione S-transferase A2	‡		3.01	
ERP29	Endoplasmic reticulum protein 29	‡	52.5	2.99	
TYB9*	Thymosin β9	‡		2.59	
CYP19A*	Cytochrome P450 19A3	‡	139	2.52	
NR5A2*	Nuclear receptor subfamily 5, group A, member 2	‡	55.1	2.45	
GART	Phosphoribosylglycinamide formyltransferase	‡		2.44	
AKR1C4*	Aldo–keto reductase family 1, member C4	‡		2.29	
AKR1C3	Aldo–keto reductase family 1, member C3	‡		2.26	
BX926910.1.p.sc.3*		‡	43.5	2.21	
HSPE1*	Heat shock 10 kDa protein 1	‡		1.98	
TFPI2	Tissue factor pathway inhibitor 2	‡		1.92	
DDX3X	DEAD (Asp–Glu–Ala–Asp) box polypeptide 3	‡		1.9	
CFL2	Cofilin 2	‡		1.84	
HNRPU	Heterogeneous nuclear ribonucleoprotein U	‡		1.82	
RFT1	Putative endoplasmic reticulum multispan transmembrane protein	‡		1.77	
BQ597365.1.p.sc.3		‡		1.56	
PSMD12	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 12	‡		1.53	
CYB5	Cytochrome b-5	‡		1.52	
CCT1	t-Complex 1	‡		1.51	

HUGO name symbol column: *identifies genes with several clones present in the selection. FDR-adjusted *P* value column: ‡$P < 0.002$. Fold change corresponds to the higher expression of the three classes versus the lowest.

Table 2 Summary of significant under-expressed genes during pig follicular development.

HUGO name symbol	Gene description	FDR-adjusted *P* value	RF importance	Maximum fold change	Expression profile
TUBA1B	Tubulin, α-1b	‡	195	2.16	
CALU	Calumenin	‡	46.5	1.98	
CB287006.1.p.sc.3		‡		1.80	
SMTN	Smoothelin	‡	80.5	1.72	
SLC40A1	Solute carrier family 40 (iron-regulated transporter), member 1	‡		1.67	
PKM2	Pyruvate kinase, muscle	‡	73.2	1.63	
CALM1, CALM2	Calmodulin 1	‡	80.1	1.51	
MT-CO1	Mitochondrially encoded cytochrome c oxidase I	*	43.7	1.61*	
CFL1	Cofilin 1	*	40.7	1.39*	
GPRC5C	G protein-coupled receptor, family C, group 5, member C	*	38	1.37*	
SFT2D2	SFT2 domain containing 2	NS	38	1.37*	
FOLR2	Folate receptor 2	†	36.1	1.35*	
GSTO1	Glutathione S-transferase omega 1	*	45.3	1.29*	
RPLP1*	Ribosomal protein, large, P1	‡		2.51	
RPS26*	Ribosomal protein S26	‡	63.2	2.47	
RPS17*	Ribosomal protein S17	‡		2.37	
EEF1A*	Eukaryotic translation elongation factor 1-α1	‡	168	2.34	
RPLP0*	Ribosomal protein, large, P0	‡	38.3	2.20	
RPL37A	Ribosomal protein L37a	‡	137	2.14	
HMGB1	High-mobility group box 1	‡		2.11	
ARL4C	ADP-ribosylation factor-like 4C	‡		2.06	
DAG1	Dystroglycan 1	‡		1.98	
TUBB5	Tubulin, β5	‡		1.97	
GNB2L1	Guanine nucleotide-binding protein (G protein), β-polypeptide 2-like 1	‡		1.94	
TB7*	Tubulin β-7 chain	‡	37.7	1.86	
ITM2A*	Integral membrane protein 2A	‡	38	1.85	
CF179049.1.p.sc.3*		‡	40.2	1.80	
RPS5	Ribosomal protein S5	‡		1.73	
H2AFZ*	H2A histone family, member Z	‡		1.62	
RPS6	Ribosomal protein S6	‡		1.50	
VIM*	Vimentin	‡		2.81	
CAPNS1	Calpain, small subunit 1	‡		2.50	
RPS25*	Ribosomal protein S25	‡		2.43	
SOX4	SRY (sex determining region Y)-box 4	‡		2.39	
RPL11	Ribosomal protein L11	‡		2.36	
STMN1*	Stathmin 1/oncoprotein 18	‡	39.2	2.33	
RPS12*	Ribosomal protein S12	‡		2.23	
RPSA*	Laminin receptor 1	‡		2.21	
BTG2	β-Cell translocation gene 2	‡		2.03	
TMSB10*	Thymosin, β10	‡		2.01	
HIST1H2AC	Histone cluster 1, H2ac	‡		1.91	
GPX3*	Glutathione peroxidase 3	‡		1.89	
EGR1*	Early growth response 1	‡		1.87	
CLTB	Clathrin, light chain	‡		1.86	
RPS8	Ribosomal protein S8	‡		1.85	
BDH2	3-Hydroxybutyrate dehydrogenase, type 2	‡		1.84	
ENTPD1*	Ectonucleoside triphosphate diphosphohydrolase 1	‡		1.82	
PPARG*	PPARG*	‡		1.77	
RPL34*	Ribosomal protein L34, leukemia-associated protein	‡		1.77	
C15ORF21	Chromosome 15 open reading frame 21	‡		1.73	
RPS7	Ribosomal protein S7	‡		1.72	
PABPC1*	Poly(A)-binding protein, cytoplasmic 1	‡		1.71	
MRPL49	Mitochondrial ribosomal protein L49	‡		1.68	
RPL3	Ribosomal protein L3	‡		1.65	
IGFBP2	Insulin-like growth factor-binding protein 2	‡		1.64	
S100A11	S100 calcium-binding protein A11	‡		1.57	
CTGF	Connective tissue growth factor	‡		1.52	
RANBP1	RAN-binding protein 1	‡		1.51	
RPL9	Ribosomal protein L9	‡		1.49	
NONO	Non-POU domain containing, octamer binding	‡		1.44	



HUGO name symbol column: *identifies genes with several clones present in the selection. FDR-adjusted *P* value column: *$P<0.05$; †$P<0.01$; ‡$P<0.002$. Fold change corresponds to the higher expression out of the three classes versus the lowest: *corresponds to genes selected by RF but not with F-analysis.

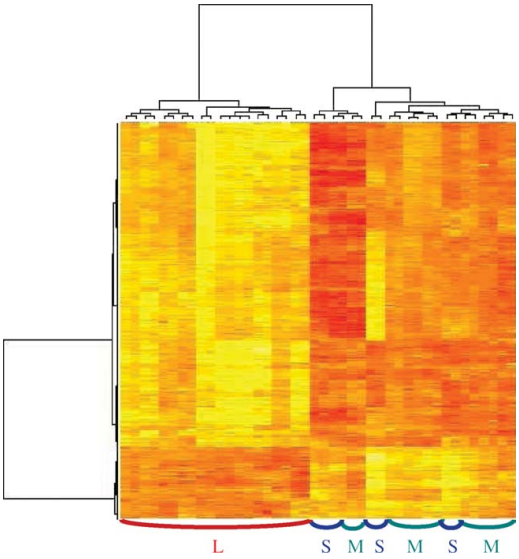


Figure 1 Heat map display of unsupervised hierarchical clustering of 643 cDNAs selected with F-analysis. The cDNAs are displayed in lines and micro-arrays in columns. The yellow colour represents over-expressed cDNAs and red down-expressed cDNAs. L corresponds to large follicles, M to medium follicles and S to small follicles.

expression of all tested genes was coherent when comparing the two approaches. Moreover, the cytochrome P450 11A (*CYP11A*) and the STAR protein (*STAR*) genes, not present on our micro-arrays and whose expression was already known to increase during the terminal development of the follicle (Hatey *et al*. 1992, Conley *et al*. 1994, LaVoie *et al*. 1997) gave the expected results.

GSTA1 gene expression, first ranked (score 222) with the RF analysis as a predictive marker for follicle size class with a relatively high expression fold change (2.58) between LF and SF, was examined by *in situ* hybridization (Fig. 4). *GSTA* mRNA was strongly detected in theca interna cells of small healthy antral follicles (∼1 mm in diameter) and not detectable in granulosa cells (Fig. 4A). By contrast, in large follicles (>5 mm), *GSTA* mRNA was strongly detected in granulosa cells (Fig. 4C).

Data integration

Grouping of differentially expressed genes into pathways was achieved by utilizing the Ingenuity Pathway Analysis (IPA). Among the 79 genes/contigs selected by F analysis, 64 genes were taken into account by IPA tool for generating networks. Five highly significant biological networks (score >16: score of 2 have at least 99% confidence of not being generated by chance) encompassing three top biological functions were identified and summarized in Table 4. The first network highlighted the down-expression of ribosomal protein genes during the follicular development as illustrated in Fig. 5. Networks 2–5 (25 focus genes) illustrated changes in the glutathione and lipid metabolism pathways (Supplementary Figure 1, which can be viewed online at www.reproduction-online.org/supplemental). Networks 3–4 (27 focus genes) were related to cellular growth and differentiation. As part of this latter network, we have particularly focused on genes related to the control of cell shape. Using the IPA tool, a specific cell shape network has been constructed. As shown in Fig. 6, this network targeted on the potential role of the actin gene. Thus, in the way to estimate granulosa cell morphological change during terminal follicular development, we have performed actin staining with FITC-conjugated phalloidin on cryosections of pig ovaries (Fig. 7). First, the observation of different sections identified a granulosa cell structure with a huge nucleus and little cytoplasm. Then, we observed different cell shapes between granulosa cells of small–medium follicles that

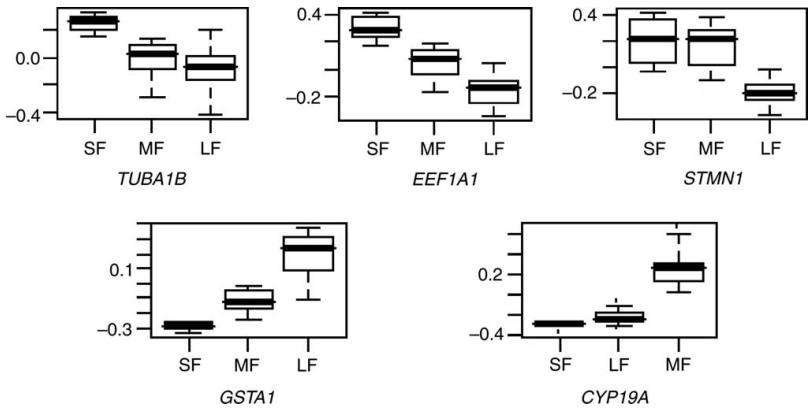


Figure 2 The different gene expression profiles. Y, log scale of gene expression.

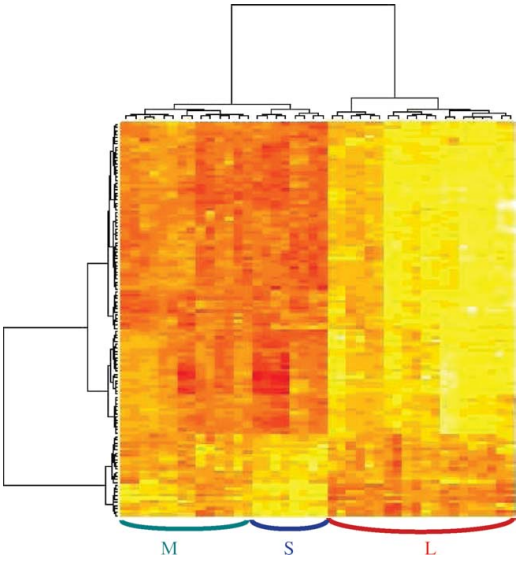


Figure 3 Heat map display of unsupervised hierarchical clustering of 120 cDNAs selected with RFs.

displayed well-defined cell shape and constituted a regular honeycomb network and granulosa cells of large ones, these had less defined borders and an elongated shape. These observations suggested a different organization of actin filaments between two types of granulosa cells. Moreover, the micro-array analysis of *actin* gene expression exhibited no differential expression.

Discussion

The objective of the present study was to identify differentially expressed genes in granulosa cells during terminal follicular development in pigs, and underline gene networks associated with this process. To reach this objective, we have developed a dedicated cDNA micro-array using SSH strategy, further hybridized with RNA probes coming from granulosa cells picked up at different steps of the terminal follicular development (SF, MF and LF).

In order to create a specific porcine micro-array, we applied SSH strategy to obtain enriched cDNA libraries with genes overexpressed in SF compared with LF and MF, or in MF and LF compared with SF. Despite the already described redundancy (Bonnet *et al.* 2006*a*), the sequence data showed a good enrichment with few overlaps between the forward and reverse libraries. For example, there is no overlap between SF/lf and LF/sf libraries, and only 11 genes/contigs (over 170) between SF/mf and MF/sf libraries. Thus, these libraries allowed the design of a valuable micro-array dedicated to the terminal follicular growth in pig species.

After micro-array hybridization and data acquisition, the statistical analyses were performed to (1) identify differentially expressed transcripts using F analysis and (2) select a set of genes that can discriminate the three follicle classes using gene prediction (RFs).

When comparing the *P* values with F analysis and T statistics, we observed a very low proportion of false positive, probably due to the pre-selection by the SSH strategy. The set of 79 selected cDNAs by F analysis

Table 3 Results of qPCR analysis.

		RT-PCR regulation			Micro-array regulation	
Way of selection	**Gene name**	Best comparison	*P* value	Fold change*	*P* value	Fold change
F/RF	BX926910.1.p.sc.3	L/S	*	328.66	‡	2.21
F/RF	NR5A2	L/S	†	5.93	‡	2.45
F/RF	GSTA	L/S	†	5.16	‡	2.58
F	HSPE1	L/S	NS	3.56	‡	1.98
RF	MT-CO1	M/S	NS	1.62	*	−1.61
F	CYB5	M/S	*	1.58	‡	1.52
F	CAPNS1	M/S	NS	−1.47	‡	−2.50
F/RF	TUBA	L/S	NS	−1.67	‡	−2.16
F	RPS5	L/S	‡	−2.15	‡	−1.73
F/RF	RPLP0	L/S	‡	−2.53	‡	−2.20
F/RF	CF179049.1.p.sc.3	L/S	*	−2.90	‡	−1.80
F	SOX4	L/M	NS	−3.25	‡	−2.39
F	HMGB1	L/M	*	−3.34	‡	−2.11
F/RF	STMN1	L/M	‡	−5.86	‡	−2.33
F	GPX3	L/M	†	−6.83	‡	−1.89
F/RF	SMTN	L/S	*	−7.42	‡	−1.72
F/RF	ITM2A	L/M	†	−16.06	‡	−1.85
F	PPARG	L/S	†	−20.04	‡	−1.77
Control +	CYP11A1	L/S	‡	14.02		
Control +	STAR	L/S	†	70.05		

Fold change corresponds to the highest expression by the lowest in relation to the best comparison. As a convention, a minus sign was added for down-regulation during development. *$P<0.05$; †$P<0.01$; ‡$P<0.002$.

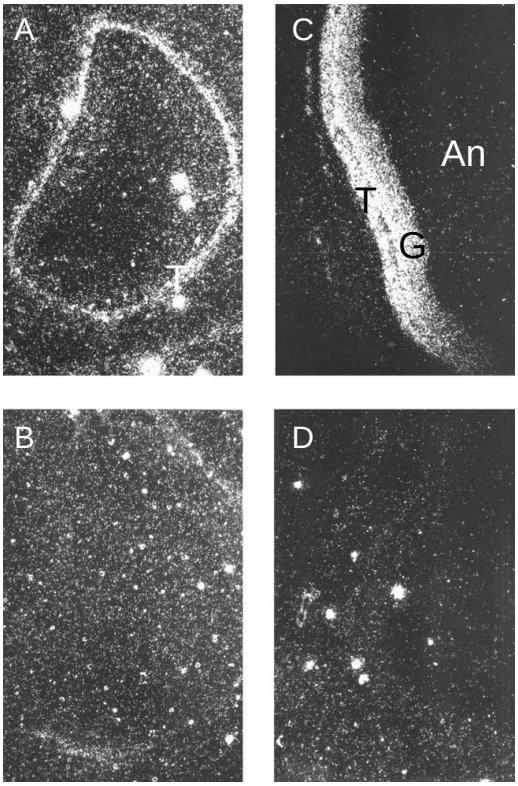


Figure 4 *In situ* hybridization of *GSTA* mRNA. Darkfield views (10× magnification) showing specific (anti-sense probe: A and C) and no specific (sense probe: B and D) hybridizations in cryosections of saw follicles at different stages of development. (A and B) Correspond to small antral follicle (<1 mm). (C and D) Correspond to large antral follicle (5 mm). T, theca cells; G, granulosa cells; An, antrum.

(FDR <0.2%) mostly favoured a hierarchical classification in two clusters, sorting out the LF from the MF and SF (Fig. 1). This observation evidences that main molecular changes occur in pig granulosa cells during the medium to large follicle transition. Interestingly, this is correlated to the appearance of functional LH receptors in granulosa cells (May & Schomberg 1984).

Despite the similarity between SF and MF classes, we tried to discriminate them using an alternative statistical method. RFs were applied to select a predictive set of genes that can classify the follicles in their respective classes (Diaz-Uriarte & Alvarez de Andres 2006, Lê Cao *et al*. 2006). The main advantages of this method are that it can deal with a massive number of correlated input variables (and hence take into account dependencies between cDNAs) and it can also select features using an internal variable importance measure. A weighting procedure is also available to deal with unbalanced classes (biological×technical replicates). RFs selected 26 genes/contigs that could predict the three classes (LF versus MF versus SF, Fig. 3). F and RF selections shared common genes (20) but they did not answer the same biological question: F analysis selected differentially expressed transcripts between one of the three follicle classes, whereas RFs selected transcripts that can predict any of the three classes preferably all together and help classify the follicles.

In order to complete these global analyses, we combined them with the individual study of each selected gene, which revealed five expression profiles (Fig. 2). These expression profiles confirmed that the main significant differences in gene expression during terminal follicular development occurred between the medium and large follicles. Indeed, 49 out of the 79 genes/contigs showed a differential expression only in large follicles (Tables 1 and 2). In addition, 23 genes/contigs exhibited a gradual expression change from SF to LF. Finally, the expression of only 13 genes/contigs decreased during SF to MF transition.

To validate micro-array analysis, gene expression was checked using real-time PCR analysis and *in situ* hybridization. The real-time PCR results were globally in agreement with the micro-array data analyses (Table 3). Discrepancies concern only five genes whose *P* values were considered as non-significant. Among them, only *MT-CO1* displays an opposite regulation between the two approaches, whereas the other four exhibited the same tendency as the one found in micro-array experiments. Non-significance may be attributed to the high variability of gene expression checked by QPCR. A higher number of tested samples would probably confirm the micro-array data for these four genes.

Using IPA, we identified major networks involved in ribosomal protein synthesis, lipid metabolism and cellular growth and proliferation (Table 4). The first network brought together the genes coding for ribosomal proteins and one translational elongation factor (*EEF1A1*), all down-expressed during terminal development (Fig. 5). This network was the most significant (score 26) and included 16 out of the 79 differentially expressed genes. It assumed a decrease in protein synthesis and may be associated with a decreased cellular growth rate. This is in agreement with previous studies describing a decrease in the percentage of proliferating granulosa cells during the final stages of follicular development in pigs and other species (Hirshfield 1986, Fricke *et al*. 1996, Pisselet *et al*. 2000). Our results suggest that the molecular mechanisms leading to granulosa cell proliferation arrest in LF occur as early as the SF to MF transition, as attested by the decreased expression of seven ribosomal genes in MF (Table 2). In addition, this network also suggested the role of the *MYC* family members known to induce ribosomal gene transcription

Table 4 Summary of functional networks for genes selected by F analysis.

Id	Genes[a]	Score	Focus genes	Top functions
1	ANXA2, CDKN2A, ↓**EEF1A1**, EPO, FOS, HRAS, IGF1R, IL3, ITGA2B, JRK, KITLG (includes EG:4254), MDM2 (includes EG:4193), MYC, MYCN, NCL, NGFB, ↓**RPL3**, ↓**RPL9**, ↓**RPL11**, ↓**RPL34**, ↓**RPL37A**, ↓**RPLP1**, RPLP2, ↓**RPLP0** (includes EG:6175), ↓**RPS5**, ↓**RPS6**, ↓**RPS7**, ↓**RPS8**, ↓**RPS12**, ↓**RPS25**, ↓**RPS26**, RPS17 (includes EG:6218), STAU1, TFAP2A, YWHAZ	26	16	Protein synthesis, cancer, cell cycle
2	AMH, ANGPTL4, AQP4, AQP7, BBC3, BCKDHA, ↑**CFL2**, ↑**CYB5A**, CYCS, ↑**DAG1**, E2F1, ↓**ENTPD1**, ↑**ERP29**, F2, ↓**GPX3**, ↓**HIST1H2AC**, ↑**HSPE1**, ↓**IGFBP2**, LAMA2, LAMB3, MAPK13, MDK, ↑**MGST1**, PAPPA, PLA2G4A, ↓**RANBP1**, ↓**RPSA**, SIM1, ↓**SOX4**, ↓**STMN1**, TF, ↑**TFPI2**, TG, THBD, TNF	24	15	Lipid metabolism, molecular transport, small molecule biochemistry
3	ABCB1B, AR, ↓**CALU** (includes EG:813), ↓**CAPNS1**, CCKAR, CIDEC, CSDE1, ↓**CTGF**, ↑**DDX3X**, ↓**EGR1**, ↓**GNB2L1**, ↑**HNRPU**, HSPH1, IFNG, MYC, NCOA4 (includes EG:8031), ↓**NONO**, ↓**PABPC1**, PCBP2, ↓**PPARG**, PRDX2, ↑**PSMC2**, PSMC4, PSMC5, PSMD3, PSMD5, PSMD7, PSMD8, PSMD9, ↑**PSMD12**, PSMD13, ↓**S100A11**, ↓**SMTN**, TERT, WT1	22	14	Cellular growth and proliferation organ development, reproductive system development and function
4	AFP, AGRIN, ↓**ARL4C**, BID, ↓**BTG2**, CCNA1, CDKN2D, CHGB, CP, CST3, ↑**CTSL**, ↑**CYP19A1**, DUSP4, ↓**EEF1A1**, ELN, F12, ↑**GART**, ↓**H2AFZ**, ↑**HADHB**, HAMP, IL6, ↓**ITM2A**, KLKB1, KRAS, LMNA, NGFB, PDIA3, ↓**RPSA**, ↓**SLC40A1**, ↑**TFPI2**, THBD, ↓**TMSB10**, TP53, TPT1, VEGF	20	13	Cellular growth and proliferation organ, genetic disorder, metabolic disease
5	ABCB11, ABCB1B, ACSL1, AFP, AKR1B7, ↑**AKR1C4**, AP3B1, AP3B2, BAAT, CETP, CLTA, ↓**CLTB**, CLTC, ↑**CYP19A1**, CYP21A2, CYP7B1, CYP8B1, ↑**GSTA1**, ↑**GSTA2**, GSTM2, GSTP1 (includes EG:2950), ↓**HMGB1**, HMGB2, HNF4A, HNF4G, ↑**HSPA8**, INS1, MTTP, ↑**NR5A2**, PDK1, ↓**PKM2**, PRLR, SP1, UGT1A9 (includes EG:54600), ↓**VIM**	14	10	Lipid metabolism, molecular transport, small molecule biochemistry

[a]Up and down arrows represent up-regulated and down-regulated genes respectively.

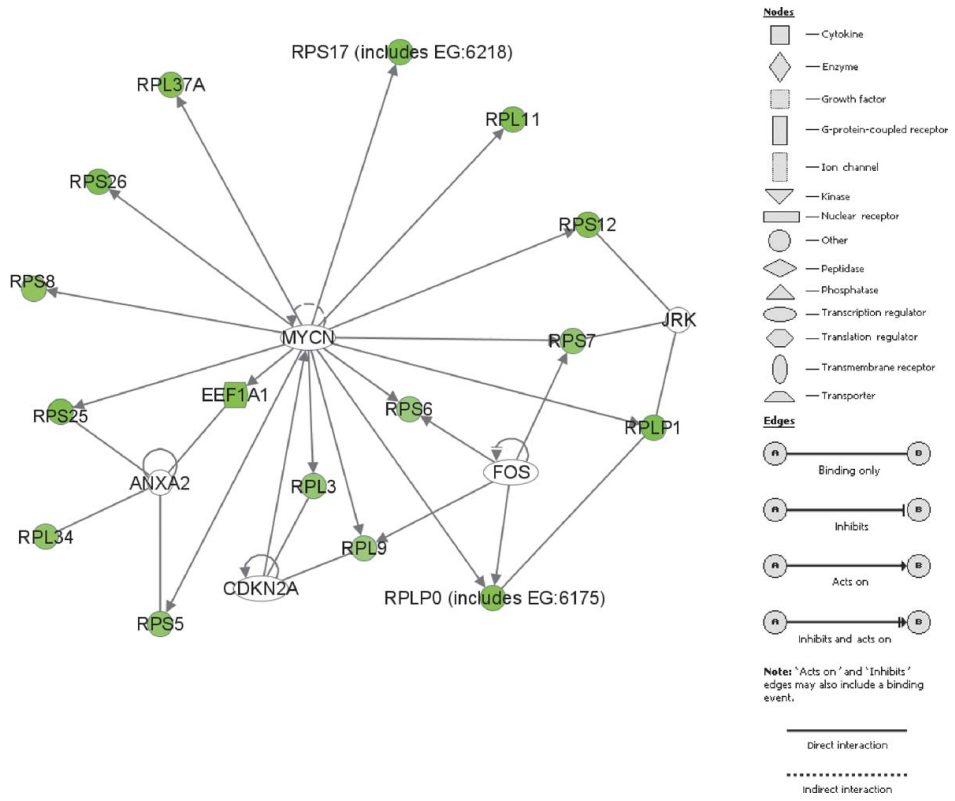


Figure 5 Network 1. Under expression of ribosomal protein genes (green colour) during follicular development.

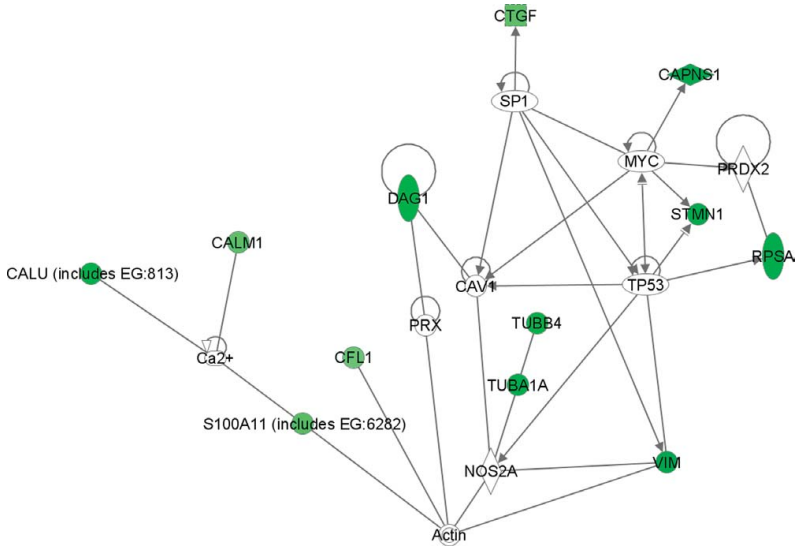


Figure 6 Granulosa cell morphology gene network.

(Boon *et al*. 2001) which needs further investigations in the context of differentiated granulosa cells.

Networks 2 and 5 bring out relations between genes implicated in lipid metabolism (Supplementary Figure 1). In the context of granulosa cells, lipid metabolism is closely related to steroids synthesis, produced *de novo* from cholesterol. We observed an overexpression of different hydroxylase genes as *HADHB,* cytochromes (*CYP19A*, *CYB5*) and *NR5A2*, a factor known to play an important role in the activation of transcription of steroidogenic enzymes like cytochromes (Sirianni *et al*. 2002) and emerging as an important ovarian factor in regulating female reproduction (Saxena *et al*. 2007, Zhao *et al*. 2007). These concomitant overexpressions occurring mainly in LF may favour an increase in steroid synthesis and were in agreement with an increase of oestradiol concentration in follicular fluid during terminal development (Foxcroft & Hunter 1985). This network also underlines a detoxification mechanism that is a consequence of steroid synthesis and allows the transformation of metabolism residues like the lipid hydroperoxydes. Our study highlighted the overexpression of different GST genes (*GSTA1*, *GSTA2* and *MGST*) in LF granulosa cells. This has already been described in steroidogenically active cells (Keira *et al*. 1994, Rabahi *et al*. 1999). In our study, *GSTA1 in situ* experiments fitted in very well with the micro-array analysis for this gene. We underlined also the down-regulation of glutathione peroxidase 3 (*GPX3*; Table 1). The KEGG (http://www.genome.jp/kegg) glutathione pathway (data not shown) suggests two different activities for these enzymes with a specific role of detoxification for GPX3 (catalyses the reduction of peroxides as lipid hydroperoxides (LOOHs)) and an intracellular transport proteins or steroids sequestration function for GSTA. To our knowledge, this is the first evidence of *GPX3* regulation in ovarian cells. Finally, the network 2 highlighted the overexpression of two members of the aldo–keto

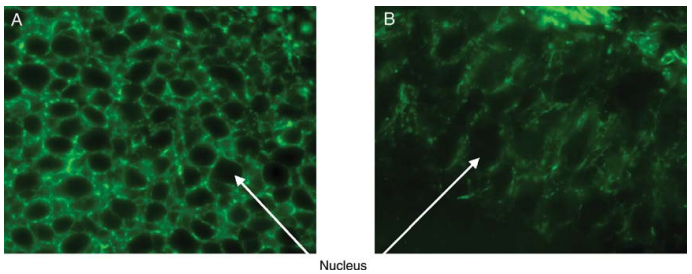


Figure 7 Cell shape. Cell shape was evaluated by actin staining with FITC-conjugated phalloidin on cryosections of pig ovaries including follicles at different stages of development. Representative microscopical field (100× magnification) of granulosa cells from (A) medium follicule, granulosa cells showed a honeycomb shape and (B) large follicle, granulosa cells showed an elongated shape.

reductase superfamily (*AKR1C3* and *AKR1C4*) able to catalyse the conversion of a wide variety of substrates such as aldehydes generated by phospholipid metabolism (Vergnes *et al*. 2003). Moreover, aldo–keto reductases have the ability to modulate the levels of active androgens, oestrogens and progestins (Bauman *et al*. 2004).

Altogether, the overexpression of the genes implicated in lipid metabolism network is in agreement with the differentiation mechanisms leading to fully steroidogenic granulosa cells in LF attested by the overexpression of *CYP19A1* (micro-array), *CYP11A1* and *STAR* genes (QPCR).

Our results notably reveal cell morphology and ion-binding gene regulation. Indeed, along terminal follicular development, we observed a down-regulation of genes coding for cytoskeletal microtubule constituents (*TUB1*, *TUB5* and *TUB7*), intermediate filaments (vimentin) and a gene implicated in their assembly, *STMN1*. This strongly suggests a deep modification of cell architecture of granulosa cells in LF. The network constructed with genes implicated in cell shape and ion binding (Fig. 6) also led us to the hypothesis of actin network implication. We showed up-regulation in LF of cofilin 2 (*CFL2*) involved in actin globular-form sequestration. In contrast, we have observed a decreased expression of *SMTN* that is specifically associated with filamentous actin in stress fibres. In addition, we observed in granulosa cells of LF the down-expression of genes coding for calcium-binding proteins, also implicated in actin cytoskeleton organization: calgizzarin (*S100A11*), *CALM1* and *CALM2* and *CALU*. Altogether, the modifications of these gene regulations let us think of the cell shape changes in LF when compared with earlier stages. This is attested by our *in vivo* observation of actin network in granulosa cells. The observation of actin staining revealed a different organization of actin filaments between the granulosa cells of small/medium and large follicles (Fig. 7). Rounded versus elongated shape may be associated with the *in vivo* differences in morphological cell gene expression between granulosa cells of SF versus LF. Interestingly, in sheep, induced cell rounding was associated with enhanced oestradiol secretion and inhibited proliferation of granulosa cells (Le Bellego *et al*. 2005). In addition, the different studies on luteinization process reported that gonadotrophic regulation of granulosa cell steroidogenesis was associated with cell shape changes (Ben-Ze'ev & Amsterdam 1989). In response to gonadotrophins, several genes coding for microtubule system and intermediate filament have been modulated. This suggests that rearrangement of the cytoskeletal proteins permits better coupling between steroidogenesis involved organelles (Carnegie *et al*. 1988, Sasson *et al*. 2004). Our hypothesis is that one of the microtubule roles tends to facilitate the movement of cholesterol from lipid droplets to mitochondria, possibly by bringing these cellular inclusions closer together (Carnegie *et al*. 1987).

This demonstrates the usefulness of building networks from micro-array experiments: they may highlight new genes that were not analysed in the micro-array experiment but that are relevant to the studied biological phenomenon, even if, in our case, the literature had already pointed these remodelling processes *in vitro*.

Up to now, only three transcriptome analyses on pig folliculogenesis have been performed. Two of them identified genes from the whole follicle whose ovarian expression has been changed as a result of long-term genetic selection for the component of reproduction (Caetano *et al*. 2004, Gladney *et al*. 2004). The third analysis examined gene expression in the whole follicle between preovulatory oestrogenic and luteinized follicles and corresponded to the continuity of our study (Agca *et al*. 2006).

In conclusion, the present study identified 79 regulated genes that may contribute to a better understanding of the mechanism involved in terminal antral follicular growth. Four cDNA sequences were found without significant homologies in databases. Some genes had functions that were already known and their regulation was consistent with the published literature. Some of them have been never associated with folliculogenesis, such as *STMN1*, *SMTN*, *ITM2A* and *GPX3*. Further investigations will be necessary to analyse the spatio-temporal expression pattern of these new genes and their interplay at the RNA and protein levels in the developing ovarian follicle. They may give clues to better understanding of the folliculogenesis process. Integration of the data highlighted gene networks mainly involved in ribosomal protein synthesis; steroid metabolism and cell morphology are perfectly coherent with enhanced steroidogenesis and arrest of growth rate in granulosa cells reaching their final maturation at the end of the terminal follicle development in pig.

Materials and Methods

Collection of ovaries

Oestrous cycles of gilts were synchronized by oral administration of 20 mg/day altrenogest (Regumate, Hoechst-Roussel, Paris, France) for 18 days at INRA experimental farm (animal experimentation authorization B-35-275-32). Ovaries were removed by laparotomy, 24 or 96 h after the last altrenogest feeding. For *in situ* experiment and phalloidin-FITC detection, ovaries were embedded in OCT medium (Miles Laboratories, CML, Nemours, France), frozen in liquid N_2 vapour and stored at −80 °C.

Granulosa cell isolation and RNA extraction

All antral follicles from 1 mm in diameter were isolated carefully using a binocular microscope. The diameter of each follicle was measured and follicles were classified according to

their size class as previously described for pig (Guthrie *et al.* 1993). SF (1–2 mm) and MF (3–4 mm) were recovered 24 h after the last altrenogest feeding of gilts. LF (≥5 mm) were recovered 96 h after the last altrenogest feeding. Granulosa cells were collected from individual follicle in MEM121/F12 (v/v) medium (Gasser *et al.* 1985). A small aliquot fraction was examined using Feulgen colouration to select cells only from healthy follicles (frequent mitosis, no pyknosis), as described previously (Monget *et al.* 1993, Besnard *et al.* 1996).

RNA was extracted from granulosa cells according to the technique described by Chomczynski & Sacchi (1987) with minor modifications (Hatey *et al.* 1995) using pools of granulosa cells from the same follicle size class. Finally, four independent RNA samples were obtained from small healthy follicles, five samples from medium healthy follicles and five samples from large healthy follicles. The quality of each RNA sample was checked through the Bioanalyser Agilent 2100 (Agilent Technologies, Massy, France).

Suppression subtractive hybridization (SSH)

Synthesis of cDNA and SSH were performed as described previously (Bonnet *et al.* 2006*a*), using SMART PCR cDNA synthesis kit and PCR-Select cDNA Subtraction kit respectively (each from Clontech). Briefly, SSH was performed using 300 ng cDNA generated from each RNA sample. After hybridization, the primary PCR amplification was achieved through 27–30 PCR cycles starting with 1 µl of a 23-fold diluted second hybridization reaction. The secondary PCR amplification was achieved through 11–12 PCR cycles starting with 1 µl of a tenfold diluted primary PCR amplification. Resulting SSH products were purified and concentrated, using Amicon Microcon-PCR filters (Millipore, St Quentin-en-Yvelines, France). cDNAs >500 bp were selected by gel filtration on 1 ml sepharose CL2B column, as described in pBluescript II XR Library Construction Kit (Stratagene Europe, Amsterdam, The Netherlands). The first 1 ml fraction was saved and ethanol precipitated with 20 µg glycogen and then diluted in 10 µl water. The selected PCR products were cloned using the pGEM-T Easy Vector (Promega) and electroporated into competent bacteria (DH10 alpha, Clontech).

Two forward SSH libraries were constructed, SF versus LF (SF/lf) and SF versus MF (SF/mf), and two reverse ones, LF versus SF (LF/sf) and MF versus SF (MF/sf). Each library contained 2500 cDNA clones and was respectively enriched with over-expressed genes in SF compared with the LF and MF, and MF and LF compared with SF.

Macro-arrays design and analysis

Bacterial clones of the four different SSH libraries (about 10 000 clones) have been spotted onto nylon filters to generate colony macro-arrays in order to sort out the false-positive clones and select differentially expressed candidate clones between the different follicle classes before sequencing.

Arrays were generated as described (Nguyen *et al.* 1995) and probed in duplicate, using both SMART products (from RNA of small, medium and large follicles) and the four SSH cDNA products (Sf/lf, LF/sf, SF/mf and MF/sf) labelled with α-^{33}P dCTP, as described previously (Hatey *et al.* 1992). After washing, the arrays were exposed 6 or 24 h to storage phosphor screens and scanned thereafter with a phosphor imaging system at a 50 µm resolution (Storm 840; GE Healthcare, Orsay, France).

Image quantification was performed using the XDotReader software and data analysis was performed using the BioPlot software (available at http://biopuce.insa-toulouse.fr). Briefly, after log transformation, the data were normalized by all spot average without background subtraction. The differential cDNA clone's expression is based on following software criterions: overexpressed threshold ratio, 1.5; under-expressed threshold ratio, 0.66 and Student's *t*-test (*P* value threshold, 0.25). A selection of differentially expressed cDNA clones was sequenced and spotted onto micro-array membranes.

Micro-arrays design and hybridization

The micro-arrays contained PCR products from 2849 pig cDNA clones coming from 1697 clones selected after SSH/macro-array experiments, 1056 clones of the AGENAE pig normalized multi-tissue cDNA library (Bonnet *et al.* 2008) and 96 clones used as controls. The multi-tissue library was used to allow a proper normalization of the data. PCR products were spotted in duplicate on two separate fields of the same nylon membrane (18×72 mm, Immobilon-NY+, Millipore) as described (Ferre *et al.* 2007). A detailed description of the resulting micro-array platform is available in the Gene Expression Omnibus database (www.ncbi.nlm.nih.gov/geo).

The quality of spotting and the relative amount of DNA in each spot have been controlled, using a ^{33}P-labelled oligonucleotide corresponding to a vector sequence present in all PCR products (Ferre *et al.* 2007; GEO accession number GSE5797 dataset). After stripping, the arrays were hybridized with ^{33}P-labelled complex probes synthesized from 5 µg of each RNA sample (4 SF, 5 MF and 5 LF), using SuperScript II RNAse H$^-$ reverse transcriptase (Invitrogen). Each complex probe has been hybridized on two individual membranes exposed 6 or 24 h to radioisotopic-sensitive imaging plates (BAS-2025; Fujifilm, Raytest, Courbevoie, France). The imaging plates were scanned thereafter with a phosphor imaging system at 25 µm resolution (BAS-5000, Fujifilm). Hybridization images obtained from oligonucleotide and complex probes were quantified using the semi-automated BZScan software (Lopez *et al.* 2004).

Data management

The experimental design, its implementation and the handling of data comply with MIAME standards (Brazma *et al.* 2001), and all the experimental data were managed using BASE software (Saal *et al.* 2002), adapted by SIGENAE bioinformatics platform (http://www.sigenae.org) to ensure radioactive experiments.

Data analysis

Using the oligonucleotide probe, spots with low signal values (i.e. <2× median of empty spots) were considered as mis-amplified or mis-spotted and were excluded from the

analysis. The data coming from complex probe hybridizations were logarithmically transformed and centred for each membrane. Spots with low signal value (below the average of empty spots +2 SDs) were considered as unexpressed and were excluded from the analysis. Finally, the remaining data were centred for each PCR product. Thereafter, a mixed linear model was fitted on these data, using the MIXED procedure of SAS software (SAS Institute Inc., Cary, NC, USA). Explanatory variables (follicle class, gene and interaction between these factors) were treated as fixed effects and animal, experimental variability (RNA, hybridization) and residual were treated as random effects.

The selection of the differentially expressed genes was performed using the R statistical software system (the Comprehensive R Archive National, http://www.cran.r-project.org). We tested the significance of the follicle classes on the gene expression using *F*-test followed by the Benjamini–Hochberg procedure controlling FDR for each cDNA (Benjamini & Hochberg 1995). Of note, *F*-test selects genes whose intra-class variance is differentially expressed in at least one of the classes. Thereafter, the expression value of each selected gene was compared between each follicle size class, using a Student's *t*-test (*P* value: 5%) to establish the expression profile along terminal follicular development.

A set of predictive genes was identified using the balanced RFs approach (Chen *et al.* 2004). To obtain a stable selection of genes, 1000 of the balanced RFs were launched with each of 15 000 trees and 42 variables randomly sampled as candidates at each split (default value proposed by R). The most important cDNAs that appeared the most frequently (in 90% of the forests) were selected using the Mean Decrease Gini importance measure as feature selection criterion.

Finally, the relevance of the two selections was evaluated via unsupervised hierarchical clustering using the Ward method and Euclidean distance with the R functions *hclust* and *heatmap* (Chipman *et al.* 2003).

Sequence annotation

Each cDNA sequence was compared with Refseq_rna mammalian database using the NCBI blastn program (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi). Blast results with an *e* value inferior to $1e-3$ were parsed and filtered to keep queries matching either a gene, a mRNA or a CDS and possessing at least a global coverage of 70% of the query sequence. Resulting hits were sorted out according to their closeness to the pig genome, their coverage and sequence identity. The selected cDNA sequences were submitted to the Human Genome Organization (HUGO) gene nomenclature committee, using their RefSeq IDs (http://www.genenames.org/). Then, HUGO gene symbols were used to name the genes.

Biological network and pathway analysis

IPA software (Ingenuity Systems Inc., Redwood City, CA, USA) was used to examine molecular pathways. This software combines functional annotations of our differentially expressed genes (focus genes) and the corresponding bibliographic data to generate significant signalling pathways and regulation networks.

Quantitative PCR analysis of gene expression

Total RNA (2 µg) from the same RNA samples used in micro-array experiments was reverse transcribed as described previously (Tosser-Klopp *et al.* 2001). An external standard (plant mRNA: I11a accession number Y10291) was added to each RNA sample (1 pg for 2 µg total RNA sample) before RT to allow quantification of the cDNA production. Primers were designed using 'Primer Express' software (Applied Biosystems, Courtaboeuf, France) and the intron–exon organization of porcine genes has been deduced by comparison with human genome using Iccare software (Muller *et al.* 2004). Translationally controlled tumour protein gene (*TCTP*, accession number BX667045) and *COX3* gene (cytochrome *c* oxidase subunit III, accession number CT971556) were found to be highly expressed but not regulated during follicle development in our micro-array experiment and used as internal controls. All primer sequences are given in Supplementary Table 1, which can be viewed online at www.reproduction-online.org/supplemental. Quantitative real-time PCR was performed using SYBR green fluorescence detection during amplification on an ABI Prism 7900 Sequence Detection System 2.1 (Applied Biosystems), according to the manufacturer's recommendations. Duplicates of each template (120 or 500 pg) were loaded in 384-well plates using a liquid handling robot (TECAN genesis rsp 200X8, Mannedorf/Zurich, Switzerland) with a 10 µl PCR mix SYBR green Power master mix (Applied Biosystems) and 0.5 µM forward and reverse primers (final volume of 13 µl). The PCR amplification conditions were as follows: 50 °C for 30 min, initial denaturation at 95 °C for 10 min and 40 cycles (95 °C for 15 s and 60 °C for 1 min). The last cycle was followed by a dissociation step (ramping to 95 °C). The real-time PCR amplification efficiency has been calculated for each primer pair with four 1:2 dilution points of the calibrator sample (pool of the 14 cDNA samples). After determination of the threshold cycle (Ct) for each sample, the PFAFFL method was applied to calculate the relative changes of each mRNA in each sample (Pfaffl 2001). The relative expression was normalized by the corresponding geometric average of an external control gene and two internal genes using geNorm v3.4 (Vandesompele *et al.* 2002). The significance of expression ratio was tested using *F*-test.

In situ hybridization and actin detection

Frozen ovaries recovered 24 or 96 h after the last altrenogest feeding were serially sectioned at a thickness of 10 µm with a cryostat. For *in situ* hybridization, ^{35}S-labelled cRNA probes (sense and anti-sense) were obtained from *GSTA* gene (accession number X93247 and X91711) by *in vitro* transcription of PCR products generated with the recombinant plasmid using primers containing T3 and T7 promoters at their 5′-end, as described previously (Besnard *et al.* 1996). For actin detection, cryosections were stained with FITC-conjugated phalloidin (Sigma–Aldrich), as previously described (Le Bellego *et al.* 2005), and were analysed using fluorescence microscopy.

Declaration of interest

The authors declare that there is no conflict of interest that would prejudice the impartiality of this scientific work.

Funding

This work was supported by a grant of the Toulouse Genopole Midi-Pyrénées program.

Acknowledgements

The authors thank Janine Rallières (UMR GC) for technical assistance, H Demay for supplying animals (INRA UMR SENAH), the Centre de Ressources Génotypage Sequençage (Toulouse Genopole Midi-Pyrénées) for technical support and SIGENAE for their support in informatics and statistical improvements. We are grateful to Yvon Tosser for careful reading of the manuscript.

References

Agca C, Ries JE, Kolath SJ, Kim JH, Forrester LJ, Antoniou E, Whitworth KM, Mathialagan N, Springer GK, Prather RS *et al.* 2006 Luteinization of porcine preovulatory follicles leads to systematic changes in follicular gene expression. *Reproduction* **132** 133–145.

Balasubramanian K, Lavoie HA, Garmey JC, Stocco DM & Veldhuis JD 1997 Regulation of porcine granulosa cell steroidogenic acute regulatory protein (StAR) by insulin-like growth factor I: synergism with follicle-stimulating hormone or protein kinase A agonist. *Endocrinology* **138** 433–439.

Bauman DR, Steckelbroeck S & Penning TM 2004 The roles of aldo–keto reductases in steroid hormone action. *Drug News & Perspectives* **17** 563–578.

Le Bellego F, Fabre S, Pisselet C & Monniaux D 2005 Cytoskeleton reorganization mediates alpha6beta1 integrin-associated actions of laminin on proliferation and survival, but not on steroidogenesis of ovine granulosa cells. *Reproductive Biology and Endocrinology* **3** 19.

Benjamini V & Hochberg V 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B* **57** 289–300.

Ben-Ze'ev A & Amsterdam A 1989 Regulation of cytoskeletal protein organization and expression in human granulosa cells in response to gonadotropin treatment. *Endocrinology* **124** 1033–1041.

Besnard N, Pisselet C, Monniaux D, Locatelli A, Benne F, Gasser F, Hatey F & Monget P 1996 Expression of messenger ribonucleic acids of insulin-like growth factor binding protein-2, -4, and -5 in the ovine ovary: localization and changes during growth and atresia of antral follicles. *Biology of Reproduction* **55** 1356–1367.

Bonnet A, Frappart PO, Dehais P, Tosser-Klopp G & Hatey F 2006*a* Identification of differential gene expression in *in vitro* FSH treated pig granulosa cells using suppression subtractive hybridization. *Reproductive Biology and Endocrinology* **4** 35.

Bonnet A, Lê Cao KA, San Cristobal M, Low-So G, Tosser-Klopp G & Hatey F 2006b Transcriptome of pig ovarian cells: discriminant genes involved in pig ovarian development. *Proceedings of the First European Conference on Pig Genomics,* Lodi, Italy, Abstract 39.

Bonnet A, Iannuccelli E, Hugot K, Benne F, Bonaldo MF, Soares MB, Hatey F & Tosser G 2008 A pig multi-tissue normalised cDNA library: large-scale sequencing, cluster analysis and 9K micro-array resource generation. *BMC Genomics* **9** 17.

Boon K, Caron HN, van Asperen R, Valentijn L, Hermus MC, van Sluis P, Roobeek I, Weis I, Voute PA, Schwab M *et al.* 2001 N-myc enhances the expression of a large set of genes functioning in ribosome biogenesis and protein synthesis. *EMBO Journal* **20** 1383–1393.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert - C, Aach J, Ansorge W, Ball CA, Causton HC *et al.* 2001 Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics* **29** 365–371.

Caetano AR, Johnson RK, Ford JJ & Pomp D 2004 Microarray profiling for differential gene expression in ovaries and ovarian follicles of pigs selected for increased ovulation rate. *Genetics* **168** 1529–1537.

Lê Cao KA, Bonnet A, Besse P, Robert-Granié C & San Cristobal M 2006 Feature selection with random forests for unbalanced multiclass microarray data: application in pig ovarian follicular development. *8th World Congress on Genetics Applied to Livestock Production,* Belo Horizonte, Brazil.

Carnegie JA, Dardick I & Tsang BK 1987 Microtubules and the gonadotropic regulation of granulosa cell steroidogenesis. *Endocrinology* **120** 819–828.

Carnegie JA, Byard R, Dardick I & Tsang BK 1988 Culture of granulosa cells in collagen gels: the influence of cell shape on steroidogenesis. *Biology of Reproduction* **38** 881–890.

Chan WK & Tan CH 1987 Induction of aromatase activity in porcine granulosa cells by FSH and cyclic AMP. *Endocrine Research* **13** 285–299.

Chen C, Liaw A & Breiman L 2004 Using random forest to learn imbalanced data. UC Berkeley.

Chipman H, Hastie T & Tibshirani R 2003 Clustering microarray data. *Statistical analysis of gene expression microarray data,* pp 159–199. New York: Chapman & Hall.

Chomczynski P & Sacchi N 1987 Single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction. *Analytical Biochemistry* **162** 156–159.

Clouscard-Martinato C, Mulsant P, Robic A, Bonnet A, Gasser F & Hatey F 1998 Characterization of FSH-regulated genes isolated by mRNA differential display from pig ovarian granulosa cells. *Animal Genetics* **29** 98–106.

Conley AJ, Howard HJ, Slanger WD & Ford JJ 1994 Steroidogenesis in the preovulatory porcine follicle. *Biology of Reproduction* **51** 655–661.

Diaz-Uriarte R & Alvarez de Andres S 2006 Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7** 3.

Drummond AE 2006 The role of steroids in follicular growth. *Reproductive Biology and Endocrinology* **4** 16.

Duda M 1997 The influence of FSH, LH and testosterone on steroidsecretion by two subpopulations of porcine granulosa cells. *Journal of Physiology and Pharmacology* **48** 89–96.

Ferre PJ, Liaubet L, Concordet D, SanCristobal M, Uro-Coste E, Tosser-Klopp G, Bonnet A, Toutain PL, Hatey F & Lefebvre HP 2007 Longitudinal analysis of gene expression in porcine skeletal muscle after post-injection local injury. *Pharmaceutical Research* **24** 1480–1489.

Foxcroft GR & Hunter MG 1985 Basic physiology of follicular maturation in the pig. *Journal of Reproduction and Fertility* **33** 1–19.

Fricke PM, Ford JJ, Reynolds LP & Redmer DA 1996 Growth and cellular proliferation of antral follicles throughout the follicular phase of the estrous cycle in Meishan gilts. *Biology of Reproduction* **54** 879–887.

Gasser F, Mulsant P & Gillois M 1985 Long-term multiplication of the Chinese hamster ovary (CHO) cell line in a serum-free medium. *In Vitro Cellular & Developmental Biology* **21** 588–592.

Gladney CD, Bertani GR, Johnson RK & Pomp D 2004 Evaluation of gene expression in pigs selected for enhanced reproduction using differential display PCR and human microarrays: I. Ovarian follicles. *Journal of Animal Science* **82** 17–31.

Guthrie HD, Bolt DJ & Cooper BS 1993 Changes in follicular estradiol-17 beta, progesterone and inhibin immunoactivity in healthy and atretic follicles during preovulatory maturation in the pig. *Domestic Animal Endocrinology* **10** 127–140.

Hatey F, Gasparoux JP, Mulsant P, Bonnet A & Gasser F 1992 P450scc regulation in pig granulosa cells: investigation into the mechanism of induction. *Journal of Steroid Biochemistry and Molecular Biology* **43** 869–874.

Hatey F, Mulsant P, Bonnet A, Benne F & Gasser F 1995 Protein kinase C inhibition of *in vitro* FSH-induced differentiation in pig granulosa cells. *Molecular and Cellular Endocrinology* **107** 9–16.

Hattori MA, Yoshino E, Shinohara Y, Horiuchi R & Kojima I 1995 A novel action of epidermal growth factor in rat granulosa cells: its potentiation of gonadotrophin action. *Journal of Molecular Endocrinology* **15** 283–291.

Hillier SG & Miro F 1993 Inhibin, activin, and follistatin. Potential roles in ovarian physiology. *Annals of the New York Academy of Sciences* **687** 29–38.

Hirshfield AN 1986 Patterns of [3H] thymidine incorporation differ in immature rats and mature, cycling rats. *Biology of Reproduction* **34** 229–235.

Hsueh AJ 1986 Paracrine mechanisms involved in granulosa cell differentiation. *Clinics in Endocrinology and Metabolism* **15** 117–134.

Jiang H, Whitworth KM, Bivens NJ, Ries JE, Woods RJ, Forrester LJ, Springer GK, Mathialagan N, Agca C, Prather RS *et al.* 2004 Large-scale generation and analysis of expressed sequence tags from porcine ovary. *Biology of Reproduction* **71** 1991–2002.

Keira M, Nishihira J, Ishibashi T, Tanaka T & Fujimoto S 1994 Identification of a molecular species in porcine ovarian luteal glutathione S-transferase and its hormonal regulation by pituitary gonadotropins. *Archives of Biochemistry and Biophysics* **308** 126–132.

LaVoie HA, Benoit AM, Garmey JC, Dailey RA, Wright DJ & Veldhuis JD 1997 Coordinate developmental expression of genes regulating sterol economy and cholesterol side-chain cleavage in the porcine ovary. *Biology of Reproduction* **57** 402–407.

Lopez F, Rougemont J, Loriod B, Bourgeois A, Loi L, Bertucci F, Hingamp P, Houlgatte R & Granjeaud S 2004 Feature extraction and signal processing for nylon DNA microarrays. *BMC Genomics* **5** 38.

May JV & Schomberg DW 1984 Developmental coordination of luteinizing hormone/human chorionic gonadotropin (hCG) receptors and acute hCG responsiveness in cultured and freshly harvested porcine granulosa cells. *Endocrinology* **114** 153–163.

Mazerbourg S, Bondy CA, Zhou J & Monget P 2003 The insulin-like growth factor system: a key determinant role in the growth and selection of ovarian follicles? a comparative species study *Reproduction in Domestic Animals* **38** 247–258.

Monget P, Monniaux D, Pisselet C & Durand P 1993 Changes in insulin-like growth factor-I (IGF-I), IGF-II, and their binding proteins during growth and atresia of ovine ovarian follicles. *Endocrinology* **132** 1438–1446.

Monget P, Fabre S, Mulsant P, Lecerf F, Elsen JM, Mazerbourg S, Pisselet C & Monniaux D 2002 Regulation of ovarian folliculogenesis by IGF and BMP system in domestic animals. *Domestic Animal Endocrinology* **23** 139–154.

Muller C, Denis M, Gentzbittel L & Faraut T 2004 The Iccare web server: an attempt to merge sequence and mapping information for plant and animal species. *Nucleic Acids Research* **32** W429–W434.

Nguyen C, Rocha D, Granjeaud S, Baldit M, Bernard K, Naquet P & Jordan BR 1995 Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones. *Genomics* **29** 207–216.

Pfaffl MW 2001 A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research* **29** e45.

Pisselet C, Clement F & Monniaux D 2000 Fraction of proliferating cells in granulosa during terminal follicular development in high and low prolific sheep breeds. *Reproduction, Nutrition, Development* **40** 295–304.

Rabahi F, Brule S, Sirois J, Beckers JF, Silversides DW & Lussier JG 1999 High expression of bovine alpha glutathione S-transferase (GSTA1, GSTA2) subunits is mainly associated with steroidogenically active cells and regulated by gonadotropins in bovine ovarian follicles. *Endocrinology* **140** 3507–3517.

Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A & Peterson C 2002 BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biology* **3** SOFTWARE0003.

Sasson R, Rimon E, Dantes A, Cohen T, Shinder V, Land-Bracha A & Amsterdam A 2004 Gonadotrophin-induced gene regulation in human granulosa cells obtained from IVF patients. Modulation of steroidogenic genes, cytoskeletal genes and genes coding for apoptotic signalling and protein kinases. *Molecular Human Reproduction* **10** 299–311.

Saxena D, Escamilla-Hernandez R, Little-Ihrig L & Zeleznik AJ 2007 Liver receptor homolog-1 and steroidogenic factor-1 have similar actions on rat granulosa cell steroidogenesis. *Endocrinology* **148** 726–734.

Shimasaki S, Moore RK, Otsuka F & Erickson GF 2004 The bone morphogenetic protein system in mammalian reproduction. *Endocrine Reviews* **25** 72–101.

Sirianni R, Seely JB, Attia G, Stocco DM, Carr BR, Pezzi V & Rainey WE 2002 Liver receptor homologue-1 is expressed in human steroidogenic tissues and activates transcription of genes encoding steroidogenic enzymes. *Journal of Endocrinology* **174** R13–R17.

Tosser-Klopp G, Bonnet A, Yerle M & Hatey F 2001 Functional study and regional mapping of 44 hormono-regulated genes isolated from a porcine granulosa cell library. *Genetics, Selection, Evolution* **33** 69–87.

Tuggle CK, Wang Y & Couture O 2007 Advances in swine transcriptomics. *International Journal of Biological Sciences* **3** 132–152.

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A & Speleman F 2002 Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology* **3** RESEARCH0034.

Vergnes L, Phan J, Stolz A & Reue K 2003 A cluster of eight hydroxysteroid dehydrogenase genes belonging to the aldo-keto reductase supergene family on mouse chromosome 13. *Journal of Lipid Research* **44** 503–511.

Zhao H, Li Z, Cooney AJ & Lan ZJ 2007 Orphan nuclear receptor function in the ovary. *Frontiers in Bioscience* **12** 3398–3405.

Received 5 July 2007
First decision 6 September 2007
Revised manuscript received 8 April 2008
Accepted 28 April 2008

12. Etude du développement embryonnaire chez le bovin

Nous étudions ici le développement d'embryons bovins inséminés artificiellement, lors de la phase d'élongation du disque embryonnaire. Quatre stades de développement sont étudiés (st2, st3, st4 et st5), dont le dernier (st5) est facilement identifiable. Ce jeu de données est caractérisé par un schéma équilibré, mais avec très peu d'embryons par stade (5), et donc des données très bruitées.
Cette étude propose une validation complète des résultats statistiques selon les étapes suivantes :

1. *sélection* : nous avons proposé d'appliquer 3 méthodes de type *wrapper* : Random Forests, OFW+CART et OFW+SVM ainsi qu'un simple test de Fisher comme méthode filtre. Il en résulte 4 listes de 50 gènes (taille des sélections choisie de façon arbitraire) ;
2. *étude des réseaux* : les biologistes ont étudié chacune des listes grâce au logiciel Ingenuity, afin d'identifier les réseaux et les fonctions des gènes sélectionnés ;
3. *validation* : parmi les 4 listes proposées, certains gènes estimés comme pertinents biologiquement sont validés par RT-PCR (sur les mêmes embryons) ;
4. *jeu de données test* : afin de tester le caractère prédictif de ces gènes, d'autres RT-PCR sont faites sur de nouveaux embryons (test en aveugle). A l'aide de la méthode Random Forests, nous proposons de prédire leur classe.

Cet étude fera l'objet d'un article qui sera très prochainement soumis.

Molecular prediction of gastrulation stages with a small extra-embryonic gene-set

Severine Degrelle [1,4] and Kim-Anh Lê Cao[2,3,4], Christèle Robert-Granié [2]
Isabelle Hue [1]

Abstract

Discriminative and predictive markers have often been looked for in cancer research to complete or refine the morphological classification of tumours. However, they have not been looked for so far in embryo development, even though there is a need of embryo classification and proper embryos staging, specifically when they do not develop synchronously. To provide a common basis in comparative studies, we looked for molecular predictors of artificially inseminated bovine extra-embryonic stages and performed a gene profiling study using 20 bovine embryos of 4 different stages while hybridising an in-house-developed bovine array dedicated to these tissues and stages. By applying 4 discriminative methods, among them classification methods (F-test Random Forests, ofwCART and ofwSVM), we ended up with 4 lists of 50 EST which corresponded to different genes, complementary functions and different potential regulators. Out of these, 11 genes which were selected by at least two approaches were assessed for their expression in the embryos. Six of these genes showed by RT-PCR the expected expression profiles, and among these 5 appeared statistically significant. The predictive accuracy of this gene set ($n = 5$) was further challenged on a batch of naïve embryos. These genes correctly predicted the stage of 70% of naïve embryos and showed for the first time that a molecular staging of extra-embryonic tissues correlates well with a gastrulation staging of embryonic tissues. These innovative results (i) pave the way to simple embryos staging (ii) reinforce the importance of the biological events which discriminate early and late differentiation processes in bovine extra-embryonic tissues and (iii) provide a new basis to compare bovine embryos developed after artificial insemination (AI), in vitro embryo production (IVP) or somatic cell nuclear transfer (SCNT).

Keywords: multiclass classification, feature selection, wrapper methods, bioanalysis, embryos, developmental staging, estradiol, progesterone.

[3]Institut de Mathématiques, Université de Toulouse et CNRS (UMR 5219), F-31062 Toulouse, France
[2]Station d'Amélioration Génétique des Animaux UR 631, Institut National de la Recherche Agronomique, F-31326 Castanet, France
[1]UMR 1198 Biologie du Developpement et Reproduction, INRA and ENVA, F-78350 Jouy en Josas and CNRS, FRE 2857, F-78350 Jouy en Josas, France
[4]These authors contributed equally to this work.

1 Introduction

In the real-world of microarray data analysis, supervised classification is still rarely used, probably because the distinction between differential and discriminative genes is not biologically obvious. Statistically, the difference between the two terms lies in the method that is performed. On one hand, seeking for differentially expressed genes belong to the inference step, where statistical hypothesis are tested. On the other hand, analytical methods (supervised or unsupervised classification) look for discriminative genes that help classifying the samples into their predefined class (in the case of supervised classification) and eventually in hidden classes (Bing et al., 2005). Both types of approaches lead to different results, as the most differentially expressed genes may not necessarily give the best predictive accuracy and hence the best discriminative genes (Allison et al., 2006; Lê Cao et al., 2007b; Bonnet et al., 2008), and vice versa.
In this study, we will make a clear distinction between *discriminative* and *predictive* genes. Discriminative means "capable of making fine distinctions" whereas predictive means "state, tell about, or make known in advance, especially on the basis of special knowledge" (Thesaurus Dictionary). As such, predictive also means that both training and test data sets are needed.

Discriminative and predictive markers have often been looked for in cancer research, due to an increasing interest in molecular classification of tumours to complete or refine previous morphological classifications. Indeed, gene expression profiles helped discriminating tumours of similar histopathological features, although varying in clinical course and/or response to treatment (DeRisi et al., 1996; Perou et al., 1999; Golub et al., 1999). To further assess the clinical utility of such discriminative genes, seen then as potential bio markers for clinical studies, in silico studies have been made to sort out the protein products localizing at the extra-cellular space or the plasma membrane (Liu et al., 2005). However, very few checked their biological validity on new tumours or biopsies (Balko et al., 2006; Hess et al., 2006). In most cases, the training and test samples were 2 partitions of the same data set or data sets previously reported.
In embryo development, classification is also needed (i) to assess development by morphology rather than chronology, especially when embryos do not develop synchronously or (ii) to provide a common basis for comparing different studies, laboratories or animal facilities over the world. Famous classifications have thus been inherited from pioneering embryologists such as Nieuwkoop and Faber for Xenopus laevis[1], Hamburger and Hamilton (1951) for Gallus gallus or Theiler (1989) but also Downs (1993) for Mus musculus[2]. Whole-mount in situ hybridisation with key molecular markers were

[1] `http://www.xenbase.org/xenbase/original/atlas/NF/NF-all.html`
[2] `http://genex.hgu.mrc.ac.uk/Databases/Anatomy/MAstaging.shtml`

further added to refine morphological classifications or to allow cross species analyses (Khakha et al., 2002) together with time-lapse imaging (Lopez-Sanchez, 2005), ultrasound imaging (Schellpfeffer et al., 2007) or cell lineage studies (C. elegans[3]), depending on the developmental model of interest. Molecular staging has however been rare, which is either based on protein maps (Latham et al., 1992) or gene expression signatures (Hamatani et al., 2004; Wang et al., 2004) of early pre-implantation stages in the mouse, or as reported recently (Mitiku and Baker, 2007) of gastrulation and organogenesis stages. In these reports, however, molecular profiling has not been used so far as a search for discriminative and/or predictive tools but as a way to understand more deeply what development means. And so it is in other mammals.

A developmental staging has indeed been developed in bovine embryos since Chang (1952); Greenstein et al. (1958) and extended to staging through molecular markers (Maddox-Hyttel et al., 2003; Hue, 2001). These markers however always relate to gastrulating stages and none to extra-embryonic staging. Nevertheless, the staging of the extra-embryonic tissues are of high interest in this species, as well as in sheep or pig, since they tremendously elongate while gastrulation proceeds and are subsequently described as sequential but distinct stages namely: ovoid, tubular and filamentous stages (Betteridge and Flechon, 1988). Molecular profiles have recently been described for these stages and species (pig: Blomberg et al., 2005, 2006, sheep: Cammas et al., 2005, cow: Ushizawa et al., 2004; Degrelle et al., 2005; Hue et al., 2007), to gain insights into a developmental process which does not occur in rodents or primates. In these reports, however, molecular profiling has not been used as a search for discriminative and/or predictive tools.

To provide a common basis in comparisons between studies and between bio technologies such as in vitro embryo production (IVP) or somatic cell nuclear transfer (SCNT), we thus looked for molecular predictors of extra-embryonic stages. Since implantation delays or defaults, as well as placenta abnormalities in SCNT, have often been correlated with delays or deficiencies in the trophoblast, one of these extra-embryonic tissues, it became important to have in hands molecular predictors of extra-embryonic stages in embryos which were developed with a classical and long used biotechnology in cattle: the artificial insemination.

To this aim, we used a gene expression profiling study performed on a training set of 20 embryos (4 stages and 5 embryos per stage) with a bovine 10K array (Hue et al., 2008). A search for discriminative genes was then achieved, applying four different statistical approaches. A simple Fisher test (F-test), that belongs to the *filter* methods, and look for differential expressed genes, and three classification methods, called *wrapper* methods, that looks for subset of discriminative genes (as opposed to test each hypothesis on each gene one by one, which is performed by the filter methods). Wrapper methods are known to be computationally costly, as they wrap classification methods

[3] `http://www.wormatlas.org/handbook/anatomyintro/anatomyintro.htm`

such as Classification and Regression Trees, (CART, Breiman et al., 1984) or Support Vector Machines (SVM, Vapnik, 1999), but they bring complementary information on the experiment (Lê Cao et al., 2007b). Random Forests (Breiman, 2001), that aggregates CART, are known to perform efficiently on high dimensional data sets, and on microarray data sets (Diaz-Uriarte and Alvarez de Andres, 2006; Bonnet et al., 2008). Two variants of a meta-algorithm called Optimal Feature Weighting (OFW, Gadat and Younes, 2007) have been proposed by Lê Cao et al. (2007b), using CART or SVM (ofwCART, ofwSVM) and were shown to bring biologically relevant results. Each of these ranking approaches allowed for the selection of genes. We further combined the results that seemed to be of biological interest to conduct further experiments and showed that indeed the selected genes could predict the different development stages on a test set of 17 naïve embryos.

2 Material and method

2.1 Embryo collection

Estrus-synchronized and inseminated Holstein cows were slaughtered on Days 19, 20, 24 and 25 of gestation (the day of artificial insemination was designated as day 0). Extra embryonic membranes and foetuses were collected. On Days 14, 15, 16, 17, and 18 of gestation embryos were non-surgically collected from super ovulated cows (AI embryos). For each conceptus, the embryonic disc was staged according to Hue (2001) whereas the extra-embryonic tissues were snap frozen in liquid nitrogen and stored at -80°C until RNA extraction.

2.2 RT-PCR

Total RNA was extracted from the extra-embryonic tissues of each embryonic cohort (array: n=20, new: n=17) with RNeasy Mini Kit according to manufacturer instructions (Qiagen). The reverse transcriptions were done on 1μg of total RNA, using 200U of Superscript II (Invitrogen) and 500ng of oligo(dT). The 1st cDNA strand was diluted 1:5 and submitted to PCR amplification in 50μl of a 1X PCR buffer: 0.4mM dNTP, 0.2μM of specific primer (MWG) and 1U of GoTaq Flexi DNA polymerase (Promega). One to 5μl of each RT was used in the PCR reactions: 1.25μl for DLD, CITED1, TGFβ3 and CALM1; 2.5μl for CAPZA2 and bPRP1 and 5μl for CPA3, β-actin, BF039259 and BF039481. CPA3, BF039481 and β-actin were submitted to 25 PCR cycles, CAPZA2, DLD, CITED1, TGFβ3, CALM1 and bPRP1 to 30 cycles, BF039259 and CN434538 to 40 cycles. PCR conditions were as described for β-actin (Degrelle et al., 2005), modified from Ushizawa et al. (2007) for bPRP1. For the other genes, see annealing temperatures in Table 1 and contact us for details on the primers.

Table 1: PCR conditions. In this table, EST corresponds to Genbank accession number and Gene ID to human RefSeqs (NCBI). Annealing temperatures are in °C and amplicon sizes in bp.

EST	Gene ID	Annealing temperature	Amplification size in bp
AW462257	CAPZA2	60	498
AW464660	DLD	60	306
AW464956	CPA3	60	326
AW465609	CITED1	60	330
BF039259	none	60	278
BF039481	none	60	310
BF042575	TGFβ3	60	301
BF043886	CALM1	60	385
CN434538	none	58	191
$\beta-$actin	AY141970	60	319
bPRP1	J02944	60	532

Table 2: EST selection with F-test, RF, CART and SVM for further analyses. EST are defined as in Table 1. The EST considered as internal control ([1]) will not be further challenged.

EST	Gene ID	F-test	RF	ofwCART	ofwSVM
BF039481	none		x	x	x
AW462257	CAPZA2	x			x
AW464956	CPA3	x	x		
CN434538	none		x	x	x
AW465609	CITED1		x	x	
BF042575	TGFβ3		x	x	x
BF043886	CALM1		x		x
AW464660	DLD	x	x	x	
BF039259	none		x	x	
bPRP1[1]	AY141970				
$\beta-$actin [1]	J02944				

[1]Internal Control

2.3 T7 linear amplification

Total RNA from AI (n=20) embryos was isolated using Trizol (Invitrogen) according to the manufacturers instructions. We used linearly amplified antisense RNA (aRNA) to hybridize the 10K array. Since the amount of mRNA was limited from in vivo conceptuses of Days 15, 16, we amplified all samples to eliminate the technical variation and to efficiently compare the gene expression data. We used MessageAmp aRNA kit (Ambion) starting from 1μg of total RNA according to Degrelle et al. (2008). aRNA purification was performed on Mini Quick Spin RNA columns (Roche Diagnostic).

2.4 The 10K nylon membrane

The 10K custom cDNA membrane developed in our laboratory was used for the experiment (Hue et al., 2008). Briefly, the array contained 7,800 EST from term placenta (Everts et al., 2005) and 2,400 EST from a new cDNA library established in collaboration with 2 other laboratories (H. Lewin et al., Urbana, Illinois, USA and J. Yang et al., Storrs, Connecticut, USA) and indexed in Unigene as Lib. 17188 and Lib. 15992 in Unigene[4]. 10,214 unique cDNA were thus spotted onto Nylon N+ membranes (Amersham Biosciences) with a 3x3 pattern (QBot, Proteigene) at the CRB GADIE platform[5]. Internal controls (30) were also included in the cDNA array.

2.5 cDNA array hybridization, image acquisition and quantification

Membrane hybridization procedures were as described by Degrelle et al. (2005, 2008). Briefly, 500ng of amplified aRNA were reverse-transcribed in the presence of [α-33P]dATP by using a Superscript II reverse transcriptase (Invitrogen). The reaction mixture was incubated for 50 min at 42°C. The labelled probes were purified on G50 Sephadex columns and diluted in 10ml of ExpressHyb Hybridization Solution (Clontech) containing 20μg of poly-A and 20μg of bovine Cot1. After a 16 hours of incubation at 68°C, the membranes were washed four times in 2X SSC/1% SDS and once in 0.1X SSC/0.5% SDS. They were then exposed to phosphor-screens for 7 days. The hybridization signals were quantified with the Imagene 5.5 software (BioDiscovery) on the PICT platform[6].

2.6 Gene expression data analysis

Pre-processing. The raw data set with 10,214 EST was then mean centered on the macroarrays to remove individual variability and log transformed. Each EST was then centered.

[4] http://www.ncbi.nlm.nih.gov
[5] http://www-crb.jouy.inra.fr/
[6] http://www.jouy.inra.fr/jouy_eng/ressources_scientifiques/transcriptomique

Application of a filter method. We applied a basic F-test that seeks if at least the variance of one group is significantly different from the other variance groups, with a False Discovery Rate controlling procedure (FDR, Benjamini and Hochberg, 1995).

Application of several wrapper methods. Classification methods reveal unstable results when the number of variables (genes or EST) is too large compared to the number of samples (macroarrays). Hence pre-filtering is advised (Dudoit et al., 2002) to infer reliable results. We hence pre-selected 2000 EST with a F-test p-value < 0.1.

Random Forests (RF, Breiman, 2001) is a well known classification and variable selection method that aggregates Classification and Regression Trees (Breiman et al., 1984). It is however constructed on bootstrap samples that can give unstable results if the number of samples per class is very small. We hence applied a stabilized version of RF (Bonnet et al., 2008). RF outputs an importance measure of all variables based on the way each tree in the forest is constructed and how important is the role of the variables splitting each node of the tree to infer a good classification of the samples (Mean Decrease Gini measure).

Optimal Feature Weigthing algorithm (OFW,Gadat and Younes, 2007), has been specifically developed for highly dimensional data like macroarrays, to deal with more than 2 classes (Lê Cao et al., 2007a). This meta-algorithm based on a strong theoretical background consists in learning the weight probability distribution on the whole set of genes based on their discriminative power. The main principle in OFW algorithm consists in repeatedly selecting a small subset of variables (genes or EST) and evaluating their predictive ability to rightly classify the macroarrays into their respective class (here the development stages). This evaluation step is performed via a classification method, such as CART or SVM, that outputs a classification error rate based on the subset of variables and on a bootstrap sample of the original macroarrays. Hence, at iteration n of the algorithm, important weights p_i^n will be given to very discriminative genes, and weights close to zero will be given to irrelevant genes (noisy or uninformative for the classification task), $i = 1 \dots G$, where G is the total number of genes which are spotted on the macroarray. Weights of all variables are then normalized so that $\sum_i p_i^n = 1$ and $\forall i \quad p_i^n \geq 0$. The next subset of genes to evaluate is then sampled with respect to this weight probability p^n. Of course the evaluation of every possible subset of variables is computationally infeasible and OFW uses stochastic approximations to overcome this problem. Note that at iteration 0, the probability weight is set to the uniform distribution (all variables can be potentially chosen at the first iteration). Both versions of OFW (ofwCART and ofwSVM, Lê Cao et al., 2007b) were applied to the data set and output a ranked list of all variables.

2.7 PCR data analysis (discriminative and predictive genes)

The embryos used for the macroarray experiments were further used to validate the discriminative ability of the selected EST. The experimental design was based on 2 RT and 1 PCR per RT and 2 gels per PCR. A simple linear model with heterogeneous variances was performed to test the effects of the genes, the development stage, the experiment type (RT-PCR1, RT-PCR-2, deposit 1 and deposit 2) and the interactions experiment $\star$ gene and experiment $\star$ stage:
$y_{ijk} = \text{gene}_i + \text{stage}_j + \text{experiment}_k + (\text{experiment} \star \text{gene})_{ik} + (\text{experiment} \star \text{stage})_{jk} + \text{e}_{ijk}$, where $\text{e}_{ijk} \sim \mathcal{N}(0, \sigma_k^2)$ and y_{ijk} is the expression of the EST i for the stage j and the experiment type k.
In addition the stage and some genes effects that were, as expected, found significant, the experiment effect as well as some interactions experiments $\star$ genes were also found significant since an amplification variation exists due to the reverse transcription or the amplification step. Indeed, the estimated variance was much lower in the two RT-PCR experiments (14,322 and 17,176) than in the deposits (31,779 and 26,186). There was in fact a significant difference between the RT-PCR and the deposits (p-values $<$0.004 to 0.01). For each gene, the PCR data was hence normalized to remove the experiment $\star$ genes and the experiment effects so that only the RT-PCR results were kept for the analysis.

2.8 Biological network and pathway analysis

The bovine EST identifiers were converted into Human RefSeq sequences and loaded into the Ingenuity Pathway Analysis software (IPA[7]). The bovine expression data were converted into significant signalling pathways and regulation networks, based on the human, mouse and rat database created by IPA which integrates both functional annotations and bibliographic data of a list of selected genes (called "focus genes"). This allowed us to assess the biological relevancy of each list of selected genes by giving a global but yet precise picture. The analyses were performed in august 2008 (last IPA update).

3 Results and discussion

The aim of this study was to discriminate bovine developmental stages based on the expression of a few genes from the extra-embryonic tissues and assess whether they could predict the real stages of such embryos. The discriminative EST were selected using F-test, RF, ofwCART and ofwSVM. We then assessed if these selected EST were good predictors of embryonic stages on naïve embryos, a decisive step which has been

[7] `http://www.ingenuity.com`

Table 3: Classical staging according to morphological landmarks of gastrulation as established in chick (HH stages for Hamburger and Hamilton 1951) or mouse embryos (Theiler stages extended by Downs 1993).These stages are visible at the microscope and refer to the essential steps of gastrulation: formation of the primitive streak, neural plate, head folds and somites. The embryonic stages defined in this study were all confirmed by the Brachyury expression profile as in Hue (2001).

Bovine stages	Morphological landmarks
2	Pre-streak
3	Early streak
4	Late streak Node
5	Neural Plate
5+	Head folds Somites: 5 to 10 pairs

Table 4: Number of common EST selected with each method (out of a 50-EST list). Note that if the EST which are selected with the wrapper methods are not the most differentially expressed (few EST in common with F-test), they still remain differentially expressed.

	F-test	RF	ofwCART	ofwSVM
F-test	#	22	18	3
RF		#	25	12
ofwCART			#	5
ofwSVM				#

rarely achieved in the literature. As control, we used a classical staging procedure (Table 3) based on the embryonic tissues of the very same embryos.

3.1 Search for discriminative genes in embryonic gene expression data sets

Based on the expression profiling study of 20 embryos with a 10K array, the selection of discriminative genes was performed with the four different approaches: F-test, RF, ofwCART and ofwSVM. Since the optimal size of the selection is still unresolved for classification methods (as it is also for filter methods), we only focused on the first ranked EST in each output list, as we knew that only a small percentage of the genes might be considered for further biological validations. Four lists of 50 EST each were then thoroughly compared with a biological interpretation rather than a statistical assessment, as the very small number of samples does not allow to evaluate the performance of the gene selections.
Common EST selected with each method were very few, a fact that is not surprising given the different aims of each method and has already been observed (Bonnet et al., 2008; Lê Cao et al., 2007b). The F-test selects genes that are differentially expressed for at least one class (stage), whereas RF, ofwCART and ofwSVM (wrapper meth-

Table 5: Summary of the top functions and pathways identified by IPA for each list of genes.

IPA Analyses (August 2008)	F-test	RF	ofwCART	ofwSVM
Top functions	Cancer Cellular growth and proliferation	Cell cycle Cellular movement	Cellular development Cancer Cell death	Cell cycle Cancer Embryonic development
Top pathways	Oxidative stress	G1/S transition of the cell cycle	TGF$-\beta$ signaling PPARα/RXR activation Oxidative stress	Gene regulation by peroxisome proliferators via PPARα G1/S transition of the cell cycle
Indirect regulators	Beta-estradiol	Beta-estradiol Retinoic acid	None	Progesterone Retinoic acid

ods) select discriminative genes that help separating all classes. Furthermore, each classification method is constructed in a different manner, with the aggregation of the classifiers CART (RF, ofwCART) or SVM (ofwSVM). This led to rather different lists. Table 4 displays the number of common EST selected with each method and shows that the RF selection seems to share the most numerous variables with the F-test and the ofw approaches. On the other hand, these latter seem to lead to almost completely different lists (see Table 4).

3.2 Functional validity of these genes

To assess the validity of the genes selected by these 4 methods, we first looked at the entire gene lists and searched for relevant functions in the bovine extra-embryonic tissues. In the figures 1 to 3 we highlighted the more significant gene networks identified by the IPA software in our gene lists. IPA identified 2 top functions within each list, top functions which involved 20 to 26 genes (ofwCART: 20, ofwSVM: 24, RF: 26, F: 25). Even though these 4 methods did not identify the same genes, they identified top functions which were all related to *Cancer*, *Cell cycle* or *Cellular growth and proliferation*. Conversely, RF, ofwCART and ofwSVM identified each a specific function: *Cellular movement*, *Cellular development* or *Embryonic development* (Table 5). Similarly, a few pathways were shared by 2 selections out of 4: *Oxidative stress* (F, ofwCART), *G1/S transition of the cell cycle* (RF, ofwSVM) as well as *PPARα* related pathways (CART, ofwSVM). *TGF-β signalling*, however, was only identified by ofwCART. Moreover, indirect interactions appeared for a few genes of these selections with *beta-estradiol* (F, RF), *retinoic acid* (RF, ofwSVM) or *progesterone* (ofwSVM).

Meaning full with regards to previous reports which underlined as well the stage-specific significance of cell growth, proliferation and migration in extra-embryonic tis-

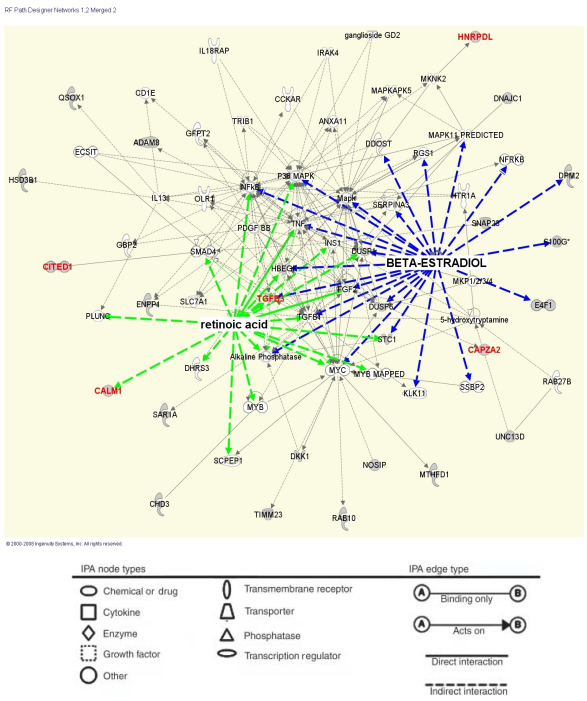


Figure 1: Merged IPA networks as of August 2008, corresponding to the 2 top functions identified with the RF selection: *Cell cycle* (38 genes) and *Cellular movement* (21 genes). Similarly to the other discriminative approaches (ofwSVM and ofwCART), RF selects discriminative genes which are not necessarily the most differentially expressed genes. Only 3 genes are common with those identified in *Cell cycle* function by the ofwSVM selection: CALM1, CAPZA2 and SNAP23 (see Fig. 4). Beta-estradiol indirectly interacts with 23 genes, 6 of which are shared by the estradiol-interacting genes from the F-test selection (see additional file 8). These genes encode DUSP1, FGF2, MYC, S100G, STC1 and TNF. All the extra-embryonic genes which are able to discriminate the embryonic stages belong to this gene set: CALM1, CAPZA2, CITED1, CPA3, HNRNPDL and TGFB3 (in red), but CPA3 did not appear in these IPA networks. As indicated in section 2, the gene ID correspond to human RefSeq. SNAP23: synaptosomal-associated protein, 23kDa, DUSP1: dual specificity phosphatase 1, FGF2: fibroblast growth factor 2 (basic), MYC: myelocytomatosis oncogene ou Nol 3, S100G: S100 calcium binding protein G, STC1: stanniocalcin 1 and TNF: tumor necrosis factor (TNF superfamily, member 2). For CALM1, CAPZA2, CITED1, HNRNPDL and TGFB3, see detailed gene ID in Table 2.

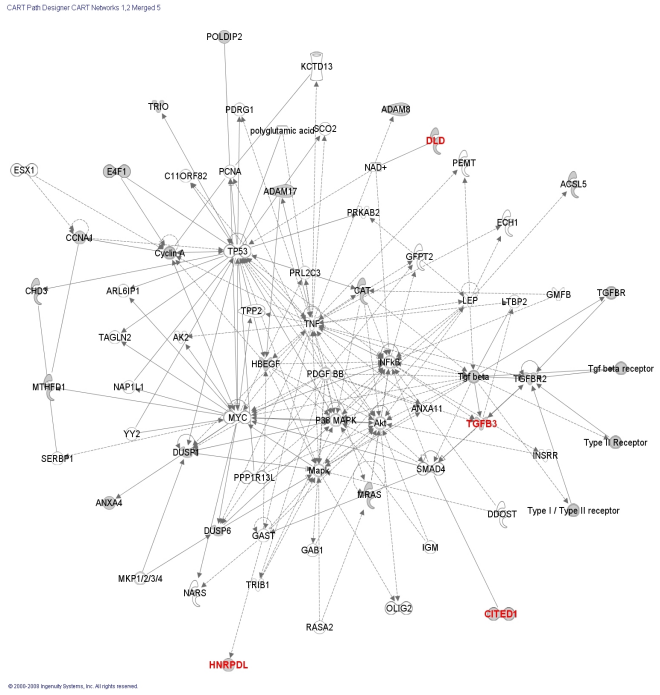


Figure 2: Merged IPA networks corresponding to the 2 top functions identified through the ofwCART selection: *Cell death* (22 genes) and *Cellular development* (22 genes). No indirect regulator appeared in the top functions shown here, but in the third one, with hydrogen peroxide. This matched with the pathways identified as oxidative stress. Hydrogen peroxide interacts there with MPV17 (MpV17 mitochondrial inner membrane protein), P53 (tumor protein p53) and TNF. Four discriminative genes out of 6 belong to this ofwCART selection: CITED1, DLD, HNRNPDL and TGFB3.

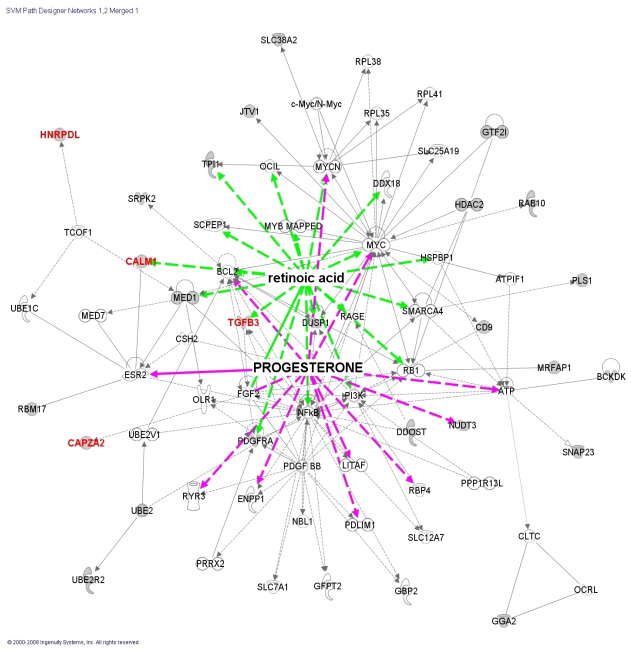


Figure 3: Merged IPA networks corresponding to the 2 top functions identified in the ofwSVM selection: *Cell cycle* (33 genes) and *Embryonic development* (22 genes). The indirect interactions appearing here involve progesterone and retinoic acid which regulate 14 and 20 genes, respectively. Five genes of this acid retinoic-interacting list were shared by the equivalent list from the RF selection: DUSP1, FGF2, MYB mapped (myeloblastosis oncogene), MYC and SCPEP1 (serine carboxypeptidase 1). Interactions with progesterone have been revealed by ofwSVM only. Four of the 6 discriminative genes emanate from this selection: CALM1, CAPZA2, HNRNPDL and TGFB3.

sues from porcine, ovine or bovine species (Blomberg et al., 2008; Cammas et al., 2005, 2006; Degrelle et al., 2005; Hue et al., 2007; Ushizawa et al., 2004; Yelich et al., 1997), these results revealed additionally indirect interactions between some of the genes and some molecules such as progesterone or estradiol, which makes sense for a developmental period occurring in utero and prior to implantation. What is new here is that these genes might be the very first bovine targets described so far on the embryonic side for these steroids. Of course, estradiol and progesterone levels in maternal blood are critical to establish the maternal support of embryo survival and development (reviewed in Spencer et al. 2004). This however has mainly been documented on the maternal side where progesterone responsive genes have been identified as critical for uterine receptivity, immune surveillance and subsequently for early embryo-uterus interactions (recently reviewed in Spencer et al. 2008; Lea and Sandra 2007, see also Bauersachs et al. 2006 and Wolf et al. 2003). In theory, the action of progesterone on the gene networks identified here may thus be mediated by an indirect effect via the uterus, a direct effect on the embryo, or possibly both. Nevertheless, progesterone receptors have been identified on the uterus (sheep: Spencer and Bazer 2002) but none so far on the embryo, neither PGR -A, -B the 2 isoforms of the PGR gene nor any of the membrane progesterone receptor components: α, β, γ. The action of estradiol is essential as well and estradiol receptors (α, β) have also been identified on the uterus (sheep: Ott et al. 1993; Spencer and Bazer 1995; pig: Ka et al. 2007). Interestingly enough, beta-estradiol is sometimes synthesized by the embryo, especially in pig where the aromatase activity is 10 times higher than in cat or roe deer and 20 times higher than in sheep or cow (Heap et al., 1981). Moreover, estradiol receptors (β) have been identified on pig embryos so that a positive loop possibly regulates the embryonic growth through an autocrine regulation (Kowalski et al., 2002). Estrogen synthesis has also been shown in elongating sheep embryos in vitro (Nephew et al., 1989), but a similar loop has not been evidenced so far. Our results might thus provide new tools to decipher such a regulation in sheep or cow.
Conversely, gene interactions with retinoic acid have already been studied in pig or sheep (Yelich et al., 1997; Cammas et al., 2006) and the networks provided here might bring new target genes, not necessarily new hypotheses.

3.3 Differential validation of discriminative genes

Among the EST that were selected by the 4 methods, we selected 11 genes that seemed discriminative for all 4 stages and were commonly selected by at least 2 statistical approaches (Table 2). These genes corresponded to 9 identified genes and 2 unidentified sequences. Due to multiple PCR amplification products (4 to 5 fragments) for one of these EST, which additionally had a high GC content, only 11 EST were further analyzed (Table 2). Among them, 2 were considered as internal controls and were not

Table 6: Expression profiles of the 6 significant EST on the embryos which were used to generate the macroarray data set (n=20).

EST	Gene ID	Methods	Expression profile	Discriminative stage
AW464660	DLD	F-RF-ofwCART	3++	3
AW464956	CPA3	F-RF	5++	5
AW465609	CITED1	RF-ofwCART	5++	5
BF039481	none	RF-ofwCART-ofwSVM	2++	2
BF042575	TGFB3	RF-ofwCART-ofwSVM	4++	4
BF043886	CALM1	ofwSVM-RF	3++	3

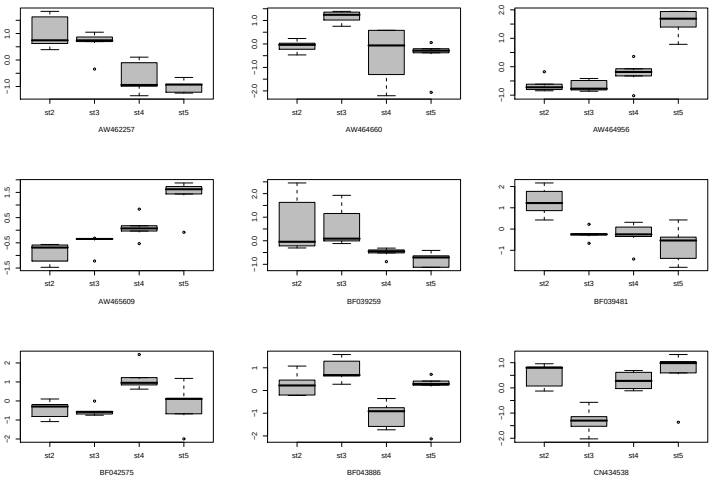


Figure 4: Expression profiles of 9 discriminative EST identified from the macroarray data set, that were analyzed with RT-PCR. Each box plot corresponds to 5 embryos per stage. Gene IDs correspond to HumRefSeq but IDs in the Bos Taurus Index from Unigene are provided too. BF039481: HNRNPDL(Heterogeneous nuclear ribonucleoprotein D-like; Bt.23277), AW462257: CAPZA2 (Capping protein (actin filament) muscle Z-line, alpha 2; Bt.13391), AW464956: CPA3 (Carboxypeptidase A3; Bt.46077), CN434538: unidentified sequence in Unigene, AW465609: CITED1 (Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 1; Bt.4437), BF042575: TGFB3 Transforming growth factor, beta 3; Bt.54513), BF043886, CALM1 (Calmodulin 1 (phosphorylase kinase, delta); Bt.61778), AW464660: DLD (Dihydrolipoamide dehydrogenase; Bt.48854), BF039259: Transcribed locus, strongly similar to NP_796352.2 ring finger protein 150 [Mus musculus]; Bt.8478).

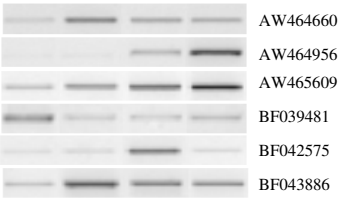


Figure 5: Illustration of the PCR results on one embryo per stage synthesized in Table 6.

further challenged. The first one (bPRP1) was known to be highly expressed at stage 5 (Ushizawa et al., 2007) and did so, the second one (β-actin) was supposed to be stable over time and was not, as it often happens with invariant housekeeping genes (see Vigneault et al. 2007; Mamo et al. 2007).
Nine EST were thus biologically analysed by RT-PCR on the embryos used to generate the array data (n=20) and a simple linear model of variance was performed on the PCR data to test the effects of the genes, the development stage, the experiment but also the interactions experiment $\star$ gene and experiment $\star$ stage (detailed in section 2.7). As a result: (i) the stage as well as some gene effects were found significant and (ii) most gene profiles, except two (DLD and β-actin), were similar between the array (Figure 4) and the RT-PCR data (Table 6 and Fig. 5).
This also happened when validating differential expression patterns using Q-PCR and has often been explained by the different sensitivity between arrays and PCR but could also be due to the different numbers of EST and/or replicates to normalize the data. Another difference may come as well from the cDNA array we used which cannot distinguish between close members of a gene family as PCR primers do. CITED1, CALM1 and CPA3 belong for example to families of 3, 4 and 6 genes respectively.

3.4 Testing the predictive validity of some discriminative genes

Since 6 EST looked properly expressed in bovine embryos as compared to their profiles in the macroarray data, they were then assessed by RT-PCR for their predictive value on a new batch of bovine embryos (n=17) collected independently from the first one. As illustrated on Figure 6, 4 EST out of 6 provided similar expression profiles by RT-PCR while 2 behaved differently, namely: BF039481 and CALM1. On this basis the 5 EST were kept to conduct a learning step with RF on the macroarray data, and a prediction step with the test data (here with the new embryos).
Keeping at the end of the selection a small number of genes was not different from similar studies in cancer research where tumours and normal tissues could be well distinguished by only 2 genes out of 2000 (colon data set), 5776 (breast data set) or 6817 (leukemia samples) as reported by Xiong et al. (2001).
Using all these genes together, we correctly predicted the embryonic stages of 13 em-

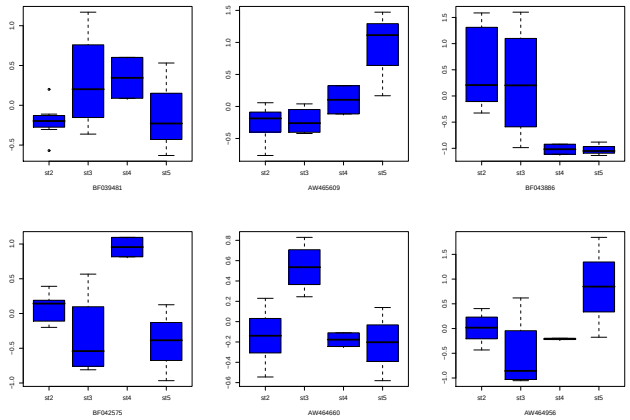


Figure 6: Expression profiles of the 6 discriminative EST on a new batch of embryos.

Table 7: Predicted stages on a new embryonic cohort (n=17), based on the 5 predictive genes identified so far. 13/17 predicted stages were correct. The problems arose for close stages, namely 2 and 3, 3 and 4.

Embryos	Predicted stages	Morphological staging	Morphological Reassessment	Brachyury WHIS
B407	5	2		
B410	5	2		
B415	2	2		
B420	2	2		
B413	2	3	2+/3-	
B416	2	3	2+/3-	
B419	2	3	3	
B424	2	3	2+/3-	
B431	4	3	3 to 4	
B432	5	3	5	
B151	2	4	3	2+/3-
B172	3	4	3	3+/4-
B175	2	4	3	3+/4-
B195	4	4		
B197	4	4		
B141	5	5		
B502	5	5		

bryos out of 17 (Table 7). Out of these 13 predictions which were good, 6 were doubtless, 7 needed re-assessment of the morphological staging and 2 were additionally confirmed by an in situ hybridisation with a well established marker of these early gastrulating stages (brachyury; Hue 2001). On the other hand, on the 4 bad predictions, only 2 were irrelevant (stage 5 predicted instead of 2). Indeed the 2 others were more subtle, predicting a stage 2 instead of a stage 3 in 2 different cases. One of these has been confirmed as wrong by in situ hybridisation, while the second wrong prediction has still to be challenged by such an in situ hybridisation. Interestingly, this pre-streak stage has already been reported as difficult to assess since morphological landmarks of gastrulation lead themselves to about 11% of wrong staging in the mouse (Downs, 1993). In cancer research, genomic predictors and clinical predictors misclassified a few cases too, as reported by Hess et al. (2006), even though these predictors were initially considered as reliable on independent replicates performed with the same array platform in 2 different laboratories. Here, we could not test our predictors on other bovine extra-embryonic data sets performed on the same array, since none was available. Even though these stages biologically overlapped the stages described here, we thus made a prediction on bovine embryos external to the training set and predicted proper stages in 70% of the naïve cohort. In similar cases, 64 to 71% of good predictions were also described for lung tumours by Balko et al. (2006).
Coming back now to embryonic staging, the only report we are aware of which dealt with staging prediction did not do so through molecular profiles but in situ hybridisation patterns on drosophila embryos (Ye et al., 2006). They also had trouble to distinguish close stages and to fix criteria for such stages. The developmental dynamics is indeed difficult to dissect into timely series of events, as evidenced in Figure 7. Indeed, the 4 stages we used to follow bovine embryo development over a 2 weeks period can be split into 4 stages but also 7 sub-stages to describe a 2 to 3 days period as detailed in the mouse embryo each 0.25 day (MGI[8]). It is hence no surprise that the stages we did not discriminate with our set of genes could correspond to late stage 2/early stage 3 or late stage 3/early stage 4. Enlarging our training set in these delicate stages or refining these stages through sub-stages as in the mouse atlas would likely improve our prediction.

3.5 Biological meaning of these predictors

Identifying predictors of embryonic stages based on molecular extra-embryonic landmarks instead of the morphological embryonic landmarks to stage gastrulating embryos was excitingly new but the question of the biological meaning of such predictors was of course challenging too. We answered them in two ways: their position in the networks revealed by IPA and their function or expression in extra-embryonic tissues, as

[8] `http://www.informatics.jax.org/`

reported in the literature.

CITED1 is a transcription factor involved in cell cycle (RF) and cell death (ofwCART) functions, that binds to other transcription factors (here SMAD4) and is normally expressed in the extra-embryonic ectoderm, the yolk sac and the trophoblast-derived cells of the placenta (Rodriguez, 2004). These expressions are recalled to Theiler stages 9 to 12 in the Mouse Gene Index[9] which corresponds nicely to the bovine stages 5 to 5+ (Table 3 and Fig. 2). Similarly, TGFB3 is increasingly expressed in the trophoblast of equivalent stages of pig embryos (Yelich et al., 1997; Gupta et al., 1998; Lee et al., 2005) though undetectable in sheep extra-embryonic tissues at similar stages (Dore et al., 1995). TGFB3 is a growth factor involved, among numerous functions, in embryonic development and epithelial to mesenchymal transition (Zavadil and Böttinger, 2005; Wyatt et al., 2007). Here, at the crossing of Cellular movement (RF) and development (ofwCART), it is preferentially expressed by bovine extra-embryonic tissues at stage 4 which is consistent with previous data. CAPZA2, which interacts here (ofwSVM) with FGF2, a growth factor involved in the positive regulation of epithelial cell proliferation, is a member of the F-actin capping protein family that regulates the growth of actin filaments and thus the actin-linked cytoskeleton organisation or biogenesis leading to cell motility. Considering the cell remodelling which occurs prior to implantation in bovine embryos, CAPZA2 could well be the visible edge of the developmental events which accompany the elongation process. CPA3, one of the 5 carboxypeptidases identified so far in human, rat or mouse, are zinc-containing exopeptidases synthesized as zymogens and activated by proteolytic cleavage. They are mainly involved in proteolysis but one of them is imprinted in human. However, CPA3 did not appear in any of the gene networks drawn by IPA. DLD is a mitochondrial enzyme involved in *Lipid metabolism* (F) as well as *Cancer and Cell death* (ofwCART). It is essential to the peri-gastrulation period in mouse embryos and covers the high metabolic needs at this time of development, especially in the mouse visceral endoderm (Johnson et al., 1997). It is described in MGI at Theiler stage 22 (day 14.5) but must be expressed earlier since the dld -/- knock out impacts on embryo development as early as E7.5, the anatomical equivalent to our stage 3 (Table 3 and Fig. 2). CALM1 might not be important by itself but through the pattern of Ca2+/calmodulin-dependent pathways and proteins it interacts with. In this study, it is related to ESR2 (estrogen receptor 2 or ER beta), but also BCL2 (B-cell CLL/lymphoma 2) and retinoic acid. Preferential expression in bovine embryos from stage 3 is however totally new to us.

Despite these biological arguments based on a gene to gene approach to shed light on their potential role *or* the developmental function, one has to keep in mind that

[9] `http://www.informatics.jax.org/`

	embryonic tissues			extra-embryonic tissues
morphological landmarks of gastrulation	**Mouse like staging Downs, 1993**	**Theiler, 1989**	Chick-like staging **Cow: Hue, 2001**	**gene expression profiles**
posterior pole	Pre-Streak 3	TS9	stage 2	
primitive streak	Early-Streak	TS9 a	stage 3	
Node (mouse) *	Late-Streak	TS10 b	stage 4	
Head-Fold - Somites	Head-Fold - Somites	TS11c - TS12b	stage 5-5+	
References	Sheep: Guillomot, 2004	MGI	Rabbit: Viebahn, 2004	

* Hensen's node (chick)

Figure 7: Alternative staging of gastrulating embryos using a small gene set from extra-embryonic tissues. Comparison to classical staging based on morphological landmarks of gastrulation.

the prediction relies on the expression pattern of the whole gene set among stages. This provides of course a more complex picture than pathways based on differentially expressed genes between 2 stages, but probably fits better to what extra-embryonic development is. Indeed, 3 cell types differentiate concomitantly while keeping constant interactions with the embryonic tissues and the maternal tissues. As such, having a small gene set with relates simultaneously to cell growth, cell differentiation, cell motility, cell signalling and to metabolism likely recalls the events which occur along extra-embryonic development so that the prediction discriminates those events (or stages) pretty well, when differing enough from each other. As for the morphological landmarks of gastrulation, where a stage is always a rather rough definition to be further declined in early, middle and late stage (HH 2-, 2, 2+ in a chick-like staging or TS10a, 10b, 10c in a mouse-like staging), it appears in this study that our molecular predictors might need as well intermediate expression levels, refined gene subsets or additional criteria to discriminate better between close biological realities.

This challenging question for the biologist can be answered in 2 to 3 ways: literature, the database or de novo in situ hybridisation. As far as CITED1, CALM1,

TGFβ3, DLD and CPA3 are concerned, we can argue with real biological data for their discriminative status. CITED1 is expressed in the mouse from Theiler stages (or TS) 9 to 12 but starts being expressed in extra-embryonic components at TS11 and is highly expressed in the yolk sac at TS12 (MGI). This corresponds nicely to the bovine stages we looked at (see Figure 7) and could account for the discriminative status of this gene. More investigation is needed regarding the other genes.

Conclusion

The top lists generated with F, RF, ofwCART and ofwSVM underlined partly known pathways and functions but revealed potential regulators which were meaningful with the biological context of a mammalian embryo developing in utero prior to implantation (estradiol, progesterone) and the literature gathered since decades on this topic in bovine, ovine and porcine species. Using several statistical approaches are thus to be recommended to look at different facets of huge expression data sets. The most classical way is to search for differential expression patterns (F-test). Here, we used 3 other methods to look for gene expression differences enabling us to discriminate among developmental stages. We successfully identified several hundreds of differences, only 200 of which were assessed. We could thus go on mining these differences to refine our predictions. On those which have been biologically (200 gene expression differences), developmentally (11) and statistically (6) assessed, it appeared in this work that RF has performed at best for the identification of stage discriminative genes (n=6/6), ofwCART and ofwSVM providing nevertheless gene selections with additional pathways or regulators. Analysing these selections together, 17 networks out of 29 were connected, leading thus to an extra-embryonic gene map involving 238 discriminative genes among which a small gene set of 5 predicted properly the developmental stage of naïve embryos with a success rate of 70%.

These discriminative genes allowed us identifying pathways or regulators of interest, such as beta-estradiol, progesterone, retinoic acid or hydrogen peroxide. One could argue that these genes are already known and did not need high throughput gene expression analyses to discover the role of these molecules. Nevertheless, this study brings new perspectives to work with: (i) potential targets of circulating progesterone on bovine extra-embryonic tissues, (ii) potential targets of circulating, or locally produced, estradiol on bovine extra-embryonic tissues. Whether these genes recall stage-specific patterns or fall into dynamic clusters of co-regulated genes is being investigated and will be addressed in another study.

These markers predicted the proper developmental stage of 70% of a naïve embryonic cohort, which is as good as what has been predicted in cancer research when array data sets were not combined with genetics or epidemiology. We thus feel confi-

dent with this very first gene set. Interestingly enough, this is the first time that one stages embryos using a small set of extra-embryonic genes instead of using molecular markers of gastrulation such as Goosecoid (Meijer et al., 2000; van den Hurk et al., 2001), Brachyury (Hue, 2001), Eomes (Guillomot et al., 2004) or Oct4 (Degrelle et al., 2005) to confirm morphological landmarks. To our knowledge, this has not be done on rabbit or mouse species.

At least but not last, this is also the first time molecular correlation is established between embryonic and extra-embryonic stages, *i.e.*, between gastrulating and elongating stages in the bovine species. This reinforces a correlation we already saw (Hue et al., 2007) and also paves the way to (i) an easy embryo staging (Fig. 6) by RT-PCR on extra-embryonic tissues, (ii) the establishment of a developmental reference to compare those stages more accurately among studies or laboratories and (iii) the refinement of predictive studies to improve the discriminative power between close stages. Whether this predictive gene set provides a basis to stage other ruminant embryos or compare bovine embryos developed after artificial insemination (AI), in vitro embryo production (IVP) or somatic cell nuclear transfer (SCNT) awaits further studies.

Acknowledgement

The authors gratefully acknowledge A. Hernandez, R. Everts, H. Lewin (University of Urbana, Illinois, USA), C. Tian and J.Yang (University of Storrs, Connecticut, USA) for their contribution to the construction of the bovine10K array and the UCEA from Bressonvilliers for the access to in vivo developed bovine conceptuses.

References

Allison, D., Cui, X., Page, G., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet, 7(1):55–65.

Balko, J., Potti, A., Saunders, C., Stromberg, A., Haura, E., and Black, E. (2006). Gene expression patterns that predict sensitivity to epidermal growth factor receptor tyrosine kinase inhibitors in lung cancer cell lines and human lung tumors. BMC Genomics, 7(1):289.

Bauersachs, S., Ulbrich, S., Gross, K., Schmidt, S., Meyer, H., Wenigerkind, H., Vermehren, M., Sinowatz, F., Blum, H., and Wolf, E. (2006). Embryo-induced transcriptome changes in bovine endometrium reveal species-specific and common molecular markers of uterine receptivity. Reproduction, 132(2):319–331.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B, 57(1):289–300.

Betteridge, K. and Flechon, J. (1988). The anatomy and physiology of pre-attachment bovine embryos. Theriogenology, 29:155–187.

Bing, N., Hoeschele, I., Ye, K., and Eilertsen, K. (2005). Finite mixture model analysis of microarray expression data on samples of uncertain biological type with application to reproductive efficiency. Veterinary Immunology and Immunopathology, 105(3-4):187–196.

Blomberg, L., Garrett, W., Guillomot, M., Miles, J., Sonstegard, T., Van Tassell, C., and Zuelke, K. (2006). Transcriptome profiling of the tubular porcine conceptus identifies the differential regulation of growth and developmentally associated genes. Molecular Reproduction and Development, 73(12):1491–1502.

Blomberg, L., Hashizume, K., and Viebahn, C. (2008). Blastocyst elongation, trophoblastic differentiation, and embryonic pattern formation. Reproduction, 135(2):181.

Blomberg, L., Long, E., Sonstegard, T., Van Tassell, C., Dobrinsky, J., and Zuelke, K. (2005). Serial analysis of gene expression during elongation of the peri-implantation porcine trophectoderm (conceptus). Physiol Genomics, 20(2):188–94.

Bonnet, A., Lê Cao, K., SanCristobal, M., Benne, F., Tosser-Klopp, G., Robert-Granié, C., Law-So, G., Besse, P., De Billy, E., Quesnel, H., et al. (2008). Identification of gene networks involved in antral follicular development. Reproduction.

Breiman, L. (2001). Random forests. Machine Learning, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA.

Cammas, L., Reinaud, P., Bordas, N., Dubois, O., Germain, G., and Charpigny, G. (2006). Developmental regulation of prostacyclin synthase and prostacyclin receptors in the ovine uterus and conceptus during the peri-implantation period. Reproduction, 131(5):917–927.

Cammas, L., Reinaud, P., Dubois, O., Bordas, N., Germain, G., and Charpigny, G. (2005). Identification of Differentially Regulated Genes During Elongation and Early Implantation in the Ovine Trophoblast Using Complementary DNA Array Screening. Biology of Reproduction, 72(4):960–967.

Chang, M. (1952). Development of bovine blastocyst with a note on implantation. Anat Rec, 113(2):143–61.

Degrelle, S., Campion, E., Cabau, C., Piumi, F., Reinaud, P., Richard, C., Renard, J., and Hue, I. (2005). Molecular evidence for a critical period in mural trophoblast development in bovine blastocysts. Developmental Biology, 288(2):448–460.

Degrelle, S., Hennequet-Antier, C., Chiapello, H., Piot-Kaminski, K., Piumi, F., Robin, S., Renard, J., and Hue, I. (2008). Amplification biases: possible differences among deviating gene expressions. BMC Genomics, 9(46).

DeRisi, J., Penland, L., Brown, P., Bittner, M., Meltzer, P., Ray, M., Chen, Y., Su, Y., and Trent, J. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet, 14(4):457–60.

Diaz-Uriarte, R. and Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7(3):1471–2105.

Dore, J., Wilkinson, J., and Godkin, J. (1995). Early Gestational Expression of Transforming Growth Factor Beta Isoforms by the Ovine Placenta. Biology of reproduction, 53:143–143.

Downs, K. (1993). Staging of gastrulating mouse embryos by morphological landmarks in the dissecting microscope. Development, 118(4):1255–1266.

Dudoit, S., Fridlyand, J., and Speed, T. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal of the American Statistical Association, 97(457):77–88.

Everts, R., Band, M., Liu, Z., Kumar, C., Liu, L., Loor, J., Oliveira, R., and Lewin, H. (2005). A 7872 cDNA microarray and its use in bovine functional genomics. Veterinary Immunology and Immunopathology, 105(3-4):235–245.

Gadat, S. and Younes, L. (2007). A stochastic algorithm for feature selection in pattern recognition. J. Mach. Learn. Res., 8:509–547.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., et al. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 286(5439):531.

Greenstein, J., Murray, R., and Foley, R. (1958). Observations on the morphogenesis and histochemistry of the bovine preattachment placenta between 16 and 33 days of gestation. Anat Rec, 132(3):321–41.

Guillomot, M., Turbe, A., Hue, I., and Renard, J. (2004). Staging of ovine embryos and expression of the T-box genes Brachyury and Eomesodermin around gastrulation. Reproduction, 127(4):491–501.

Gupta, A., Ing, N., Bazer, F., Bustamante, L., and Jaeger, L. (1998). Beta transforming growth factors (TGF beta) at the porcine conceptus-maternal interface. I. Expression of TGF beta 1, TGF beta 2, and TGF beta 3 messenger ribonucleic acids. Biology of reproduction, 59(4):905–910.

Hamatani, T., Carter, M., Sharov, A., and Ko, M. (2004). Dynamics of Global Gene Expression Changes during Mouse Preimplantation Development. Developmental Cell, 6(1):117–131.

Hamburger, V. and Hamilton, H. (1951). A series of normal stages in the development of the chick embryo. 1951. J Morphol, 88:49–92.

Heap, R., Flint, A., and Gadsby, J. (1981). Embryonic signals and maternal recognition. Cellular and Molecular Aspects of Implantation, pages 311–325.

Hess, K., Anderson, K., Symmans, W., Valero, V., Ibrahim, N., Mejia, J., Booser, D., Theriault, R., Buzdar, A., Dempsey, P., et al. (2006). Pharmacogenomic Predictor of Sensitivity to Preoperative Chemotherapy With Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide in Breast Cancer. Journal of Clinical Oncology, 24(26):4236.

Hue, I. (2001). Brachyury is expressed in gastrulating bovine embryos well ahead of implantation. Development Genes and Evolution, 211(3):157–159.

Hue, I., Degrelle, S., Campion, E., and Renard, J. (2007). Gene expression in elongating and gastrulating embryos from ruminants. Reproduction in Domestic Ruminants VI.

Hue, I. et al. (2008). Gene expression dynamics in bovine extra-embryonic tissues before placentation. Technical report, INRA Jouy-en-Josas.

Johnson, M., Yang, H., Magnuson, T., and Patel, M. (1997). Targeted disruption of the murine dihydrolipoamide dehydrogenase gene (Dld) results in perigastrulation lethality. Proceedings of the National Academy of Sciences, 94(26):14512–14517.

Ka, H., Al-Ramadan, S., Erikson, D., Johnson, G., Burghardt, R., Spencer, T., Jaeger, L., and Bazer, F. (2007). Regulation of Expression of Fibroblast Growth Factor 7 in the Pig Uterus by Progesterone and Estradiol. Biology of Reproduction, 77(1):172.

Khakha, M., Chung, C., Bustamante, E., Gaw, L., Trott, K., Yeh, J., Lim, N., Lin, J., Taverner, N., Amaya, E., et al. (2002). Techniques and Probes for the Study of Xenopus tropicalis Development. Developmental Dynamics, 225:499–510.

Kowalski, A., Graddy, L., Vale-Cruz, D., Choi, I., Katzenellenbogen, B., Simmen, F., and Simmen, R. (2002). Molecular Cloning of Porcine Estrogen Receptor-β Complementary DNAs and Developmental Expression in Periimplantation Embryos 1. Biology of Reproduction, 66(3):760–769.

Latham, K., Garrels, J., Chang, C., and Solter, D. (1992). Analysis of embryonic mouse development: construction of a high-resolution, two-dimensional gel protein database. Appl Theor Electrophor, 2(6):163–70.

Lê Cao, K.-A., Bonnet, A., and Gadat, S. (2007a). Multiclass classification and gene selection with a stochastic algorithm. Technical report, Université de Toulouse, Institut National de la Recherche Agronomique.

Lê Cao, K.-A., Gonçalves, O., Besse, P., and Gadat, S. (2007b). Selection of biologically relevant genes with a wrapper stochastic algorithm. Statistical Applications in Genetics and Molecular Biology, 6(:Iss. 1):Article 1.

Lea, R. and Sandra, O. (2007). Immunoendocrine aspects of endometrial function and implantation. Reproduction, 134(3):389–404.

Lee, S., Zhao, S., Recknor, J., Nettleton, D., Orley, S., Kang, S., Lee, B., Hwang, W., and Tuggle, C. (2005). Transcriptional profiling using a novel cDNA array identifies differential gene expression during porcine embryo elongation. Molecular Reproduction and Development, 71(2):129–139.

Liu, J., Cutler, G., Li, W., Pan, Z., Peng, S., Hoey, T., Chen, L., and Ling, X. (2005). Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics, 21(11):2691–2697.

Maddox-Hyttel, P., Alexopoulos, N., Vajta, G., Lewis, I., Rogers, P., Cann, L., Callesen, H., Tveden-Nyborg, P., and Trounson, A. (2003). Immunohistochemical and ultrastructural characterization of the initial post-hatching development of bovine embryos. Reproduction, 125(4):607–23.

Mamo, S., Gal, A., Bodo, S., and Dinnyes, A. (2007). Quantitative evaluation and selection of reference genes in mouse oocytes and embryos cultured in vivo and in vitro. BMC Developmental Biology, 7(1):14.

Meijer, H., Van De Pavert, S., Stroband, H., and Boerjan, M. (2000). Expression of the organizer specific homeobox gene Goosecoid(gsc) in porcine embryos. Molecular Reproduction and Development, 55(1):1–7.

Mitiku, N. and Baker, J. (2007). Genomic Analysis of Gastrulation and Organogenesis in the Mouse. Developmental Cell, 13(6):897–907.

Nephew, K., McClure, K., and Pope, W. (1989). Embryonic Migration Relative to Maternal Recognition of Pregnancy in Sheep. Journal of Animal Science, 67(4):999.

Ott, T., Zhou, Y., Mirando, M., Stevens, C., Harney, J., Ogle, T., and Bazer, F. (1993). Changes in progesterone and oestrogen receptor mRNA and protein during maternal recognition of pregnancy and luteolysis in ewes. Journal of Molecular Endocrinology, 10(2):171–183.

Perou, C., Jeffrey, S., van de Rijn, M., Rees, C., Eisen, M., Ross, D., Pergamenschikov, A., Williams, C., Zhu, S., Lee, J., et al. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci US A, 96(16):9212–7.

Schellpfeffer, M., Bolender, D., and Kolesari, G. (2007). High frequency ultrasound imaging of the growth and development of the normal chick embryo. Ultrasound Med Biol., 33(5):751–61.

Spencer, T. and Bazer, F. (1995). Temporal and spatial alterations in uterine estrogen receptor and progesterone receptor gene expression during the estrous cycle and early pregnancy in the ewe. Biology of Reproduction, 53(6):1527–1543.

Spencer, T. and Bazer, F. (2002). Biology of progesterone action during pregnancy recognition and maintenance of pregnancy. Front Biosci, 7:d1879–1898.

Spencer, T., Johnson, G., Bazer, F., and Burghardt, R. (2004). Implantation mechanisms: insights from the sheep. Reproduction, 128(6):657.

Spencer, T., Sandra, O., and Wolf, E. (2008). Genes involved in conceptus-endometrial interactions in ruminants: insights from reductionism and thoughts on holistic approaches. Reproduction, 135(2):165.

Theiler, K. (1989). The house mouse: atlas of embryonic development. New York: Springer-Verlag.

Ushizawa, K., Herath, C., Kaneyama, K., Shiojima, S., Hirasawa, A., Takahashi, T., Imai, K., Ochiai, K., Tokunaga, T., Tsunoda, Y., et al. (2004). cDNA microarray analysis of bovine embryo gene expression profiles during the pre-implantation period. Reprod Biol Endocrinol, 2:77.

Ushizawa, K., Takahashi, T., Hosoe, M., Kizaki, K., Abe, Y., Sasada, H., Sato, E., and Hashizume, K. (2007). Gene expression profiles of novel caprine placental prolactin-related proteins similar to bovine placental prolactin-related proteins. BMC Developmental Biology, 7(1):16.

van den Hurk, R., Taverne, M., Boerjan, M., and Stroband, H. (2001). Comparison of anterior–posterior development in the porcine versus chicken embryo, using goosecoid expression as a marker. Reprod. Fertil. Dev, 13:177–185.

Vapnik, V. N. (1999). The Nature of Statistical Learning Theory (Information Science and Statistics). Springer.

Vigneault, C., Gilbert, I., Sirard, M., and Robert, C. (2007). Using the histone H2a transcript as an endogenous standard to study relative transcript abundance during bovine early development. Mol Reprod Dev., 74(6):703–15.

Wang, Q., Piotrowska, K., Ciemerych, M., Milenkovic, L., Scott, M., Davis, R., and Zernicka-Goetz, M. (2004). A Genome-Wide Study of Gene Activity Reveals Developmental Signaling Pathways in the Preimplantation Mouse Embryo. Developmental Cell, 6(1):133–144.

Wolf, E., Arnold, G., Bauersachs, S., Beier, H., Blum, H., Einspanier, R., Frohlich, T., Herrler, A., Hiendleder, S., Kolle, S., et al. (2003). Embryo-Maternal Communication in Bovine-Strategies for Deciphering a Complex Cross-Talk. Reproduction in Domestic Animals, 38(4):276–289.

Wyatt, L., Wadham, C., Crocker, L., Lardelli, M., and Khew-Goodall, Y. (2007). The protein tyrosine phosphatase Pez regulates TGF {beta}, epithelial mesenchymal transition, and organ development. The Journal of Cell Biology, 178(7):1223.

Xiong, M., Li, W., Zhao, J., Jin, L., and Boerwinkle, E. (2001). Feature (Gene) Selection in Gene Expression-Based Tumor Classification. Molecular Genetics and Metabolism, 73(3):239–247.

Ye, J., Chen, J., Li, Q., and Kumar, S. (2006). Classification of Drosophila embryonic developmental stage range based on gene expression pattern images. Comput Syst Bioinformatics Conf, 8(293).

Yelich, J., Pomp, D., and Geisert, R. (1997). Detection of transcripts for retinoic acid receptors, retinol-binding protein, and transforming growth factors during rapid trophoblastic elongation in the porcine conceptus. Biology of Reproduction, 57(2):286–294.

Zavadil, J. and Böttinger, E. (2005). TGF-β and epithelial-to-mesenchymal transitions. Oncogene, 24:5764–5774.

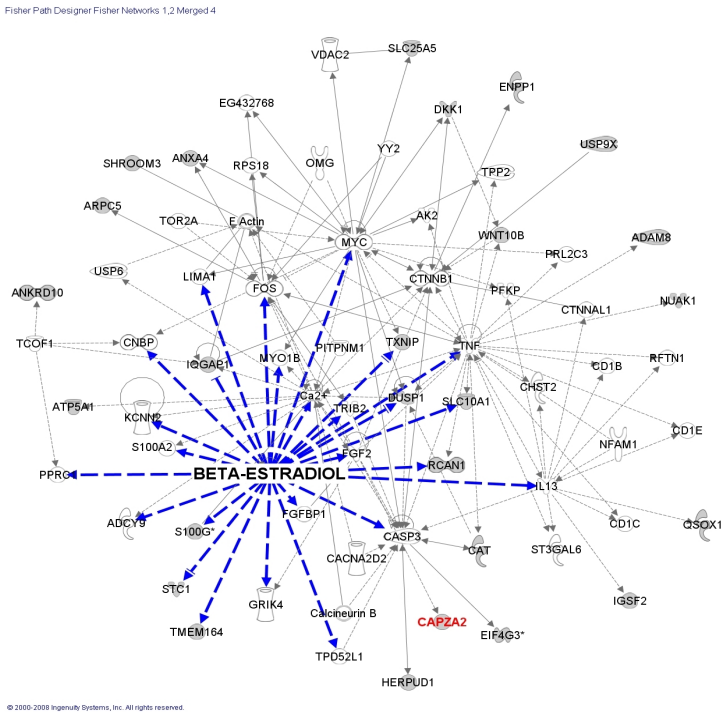


Figure 8: Supplemental figure: merged IPA networks corresponding to the 2 top functions identified through the F-test selection: *Cancer* (32 genes) and *Cellular growth and proliferation* (26 genes). The F-test is the most classical method to analyse array data sets and identify significant gene expression differences. The indirect interactions shown here involve beta-estradiol and 25 genes. Of the 6 discriminative genes, 3 emanate from this selection: CAPZA2, CPA3 and DLD, but CPA3 did not appear in these networks drawn by IPA.

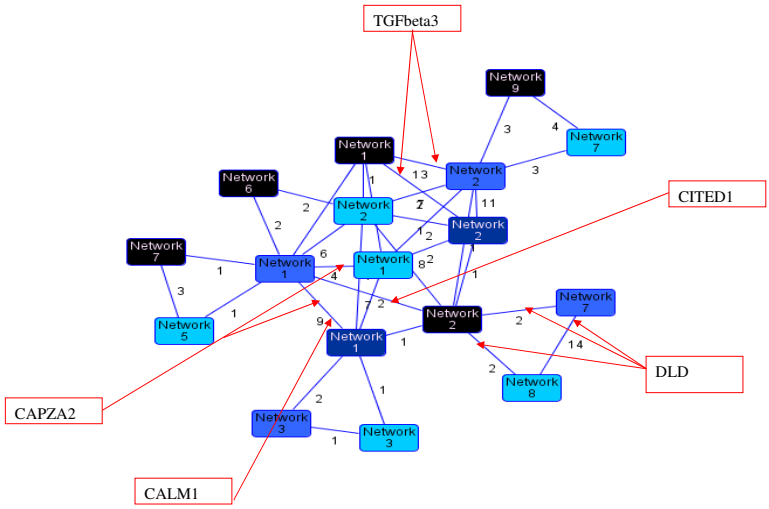


Figure 9: Supplemental figure: complementarity of the variable selection approaches, gene networks and stage-discriminative genes. Networks associated with the gene lists selected with ofwSVM (dark blue), RF (blue), F-test (light blue) and ofwCART (black) are schematically represented. CITED1 was selected with ofwCART and RF; CALM1 with ofwSVM and RF; TGBβ3 with ofwSVM, ofwCART and RF; DLD with ofwCART, F-test and RF; CAPZA2 with ofw, RF and F-test.

Bibliographie

Abdi, H. (2003). Partial least squares (PLS) regression. Encyclopedia of social sciences research methods (ed. M. Lewis–Beck, A. Bryman and T. Futing), pages 1–7.

Aha, D. & Bankert, R. (1995). A comparative evaluation of sequential feature selection algorithms. Learning from Data : Artificial Intelligence and Statistics V, pages 199–206.

Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403 :503–511.

Allison, D., Cui, X., Page, G., & Sabripour, M. (2006). Microarray data analysis : from disarray to consolidation and consensus. Nat Rev Genet, 7(1) :55–65.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., & Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America, 96(12) :6745.

Ambroise, C. & McLachlan, G. J. (2002). Selection bias in gene extraction in tumour classification on basis of microarray gene expression data. Proc. Natl. Acad. Sci. USA, 99(1) :6562–6566.

Antoniadis, A., Lambert-Lacroix, S., & Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics, 19(5) :563–570.

Aspremont, A., Bach, F., & El Ghaoui, L. (2008). Optimal Solutions for Sparse Principal Component Analysis. Arxiv preprint arXiv :0707.0705v4.

Aspremont, A., El Ghaoui, L., Jordan, M., & Lanckriet, G. R. G. (2007). A Direct Formulation for Sparse PCA Using Semidefinite Programming. Siam Review, 49(3) :434–448.

Baccini, A., Besse, P., Déjean, S., Martin, P. G., Robert-Granié, C., & SanCristobal, M. (2005). Stratégies pour l'analyse statistique de données transcriptomiques. Journal de la Société Française de Statistique, 146(1-2) :5–44.

Biau, G., Devroye, L., & Lugosi, G. (2007). Consistency of random forests and other averaging classifiers. Technical report, Université Pierre et Marie Curie, Paris.

Bjorck, A. & Golub, G. (1973). Numerical methods for computing angles between linear subspaces. Mathematics of Computation, 27(123) :579–594.

Blum, A. & Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, 97(1-2) :245–271.

Bonnet, A., Lê Cao, K., SanCristobal, M., Benne, F., Tosser-Klopp, G., Robert-Granié, C., Law-So, G., Besse, P., De Billy, E., Quesnel, H., et al. (2008). Identification of gene networks involved in antral follicular development. Reproduction.

Boulesteix, A. (2004). PLS Dimension Reduction for Classification with Microarray Data. Statistical Applications in Genetics and Molecular Biology, 3(1) :1075.

Boulesteix, A. & Strimmer, K. (2005). Predicting transcription factor activities from combined analysis of microarray and chip data : a partial least squares approach. Theor Biol Med Model, 2(23).

Braga-Neto, U. & Dougherty, E. (2004). Is cross-validation valid for small-sample microarray classification ? Bioinformatics, 20(3) :374–380.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2) :123–140.

Breiman, L. (2001). Random forests. Machine Learning, 45(1) :5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA.

Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares Jr, M., & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences, 97(1) :262.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K., Hayward, B., Keith, T., & Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests. Genetic Epidemiology, 28(2) :171–182.

Burges, C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2(2) :121–167.

Bylesjö, M., Eriksson, D., Kusano, M., Moritz, T., & Trygg, J. (2007). Data integration in plant biology : the o2pls method for combined modeling of transcript and metabolite data. The Plant Journal, 52 :1181–1191.

Cadima, J. & Jolliffe, I. (1995). Loading and correlations in the interpretation of principle compenents. Journal of Applied Statistics, 22(2) :203–214.

Candes, E. & Tao, T. (2007). The Dantzig selector : Statistical estimation when p is much larger than n. Annals of Statistics, 35(6).

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. Technical Report 666, Dpt. of Statistics, University of Berkeley.

Chessel, D. & Hanafi, M. (1996). Analyses de la co-inertie de K nuages de points. Revue de Statistique Appliquée, 44(2) :35–60.

Combes, S., González, I., Déjean, S., Baccini, A., Jehl, N., Juin, H., Cauquil, L., & Béatrice Gabinaud, François Lebas, C. L. (2008). Relationships between sensorial and physicochemical measurements in meat of rabbit from three different breeding systems using canonical correlation analysis. Meat Science, in press.

Culhane, A., Perriere, G., & Higgins, D. (2003). Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. BMC Bioinformatics, 4(1) :59.

de Jong, S. (1993). Simpls : An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 18 :251–263.

de Koning, D., Jaffrezic, F., Lund, M., Watson, M., Channing, C., Hulsegge, I., Pool, M., Buitenhuis, B., Hedegaard, J., Hornshoj, H., et al. (2007). The EADGENE Microarray Data Analysis Workshop (Open Access publication). Genet Sel Evol, 39(6) :621–31.

Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. Bioinformatics, 20(18) :3583–3593.

Dettling, M. & Buhlmann, P. (2003). Boosting for tumor classification with gene expression data. Bioinformatics, 19(9) :1061–1069.

Diaz-Uriarte, R. & Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7(3) :1471–2105.

Doledec, S. & Chessel, D. (1994). Co-inertia analysis : an alternative method for studying species-environment relationships. Freshwater Biology, 31(3) :277–294.

Donoho, D. & Johnstone, I. (1992). Ideal Spatial Adaptation via Wavelet Shrinkage. Biometrika, 81(3) :425–455.

Dray, S., Chessel, D., & Thioulouse, J. (2003). Co-inertia analysis and the linking of ecological data tables. Ecology, 84(11) :3078–3089.

Dudoit, S., Fridlyand, J., & Speed, T. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal of the American Statistical Association, 97(457) :77–88.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. Annals of Statistics, 32(2) :407–499.

Efron, B. & Tibshirani, R. (1997). Improvements on cross-validation : the e.632+ bootstrap method. Journal of American Statistical Association, 92 :548–560.

Eisen, M., Spellman, P., Brown, P., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 95(25) :14863.

Eitrich, T., Kless, A., Druska, C., Meyer, W., & Grotendorst, J. (2007). Classification of Highly Unbalanced CYP450 Data of Drugs Using Cost Sensitive Machine Learning Techniques. Journal of Chemical Information and Modeling, 47(1) :92–103.

Fan, J. & Li, R. (2001). Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties. Journal of the American Statistical Association, 96(456) :1348–1361.

Freund, Y. & Schapire, R. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences, 55(1) :119–139.

Fu, W., Carroll, R., & Wang, S. (2005). Estimating misclassification error with small samples via bootstrap cross-validation. Bioinformatics, 21(9) :1979–1986.

Gadat, S. & Younes, L. (2007). A stochastic algorithm for feature selection in pattern recognition. J. Mach. Learn. Res., 8 :509–547.

Gittins, R. (1985). Canonical Analysis : A Review with Applications in Ecology. Springer-Verlag.

Golub, G. & Van Loan, C. (1996). Matrix Computations. Johns Hopkins University Press.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., et al. (1999). Molecular Classification of Cancer : Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 286(5439) :531.

González, I. (2008). Analyse Canonique Régularisée pour des données fortement multidimensionnelles. PhD thesis, Université de Toulouse, France.

González, I., Déjean, S., Martin, P. G. P., & Baccini, A. (2008). Cca : An r package to extend canonical correlation analysis. Journal of Statistical Software, 23(12).

Guyon, I., Elisseefi, A., & Kaelbling, L. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3(7-8) :1157–1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 46(1) :389–422.

Hardoon, D., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical Correlation Analysis : An Overview with Application to Learning Methods. Neural Computation, 16(12) :2639–2664.

Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D., & Brown, P. (2000). Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol, 1(2) :1–21.

Hirai, M., Klein, M., Fujikawa, Y., Yano, M., Goodenowe, D., Yamazaki, Y., Kanaya, S., Nakamura, Y., Kitayama, M., Suzuki, H., et al. (2005). Elucidation of Gene-to-Gene and Metabolite-to-Gene Networks in Arabidopsis by Integration of Metabolomics and Transcriptomics. Journal of Biological Chemistry, 280(27) :25590.

Hirai, M., Yano, M., Goodenowe, D., Kanaya, S., Kimura, T., Awazuhara, M., Arita, M., Fujiwara, T., & Saito, K. (2004). From The Cover : Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. Proceedings of the National Academy of Sciences, 101(27) :10205.

Hoerl, A. & Kennard, R. (1984). Ridge regression in 'Encyclopedia of Statistical Sciences', volume 8. Wiley, New York, Monterey, CA.

Hoskuldsson, A. (1988). PLS regression methods. Journal of Chemometrics, 2(3) :211–228.

Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 28 :321–377.

Hsu, C. & Lin, C. (2002). A comparison of methods for multiclass support vector machines. Neural Networks, IEEE Transactions on, 13(2) :415–425.

Izmirlian, G. (2004). Application of the Random Forest Classification Algorithm to a SELDI-TOF Proteomics Study in the Setting of a Cancer Prevention Trial. Annals of the New York Academy of Sciences, 1020(1 The Applications of Bioinformatics in Cancer Detection) :154–174.

Jaffrézic, F., de Koning, D., Boettcher, P., Bonnet, A., Buitenhuis, B., Closset, R., Déjean, S., Delmas, C., Detilleux, J., Dovc, P., et al. (2007). Analysis of the real EADGENE data set : Comparison of methods and guidelines for data normalisation and selection of differentially expressed genes (Open Access publication). Genet. Sel. Evol, 39 :633–650.

James, G., Radchenko, P., & Lv, J. (2007). The dasso algorithm for fitting the dantzig selector and the lasso. Technical report, Marshall School of Business, University of Southern California.

Jeffers, J. (1967). Two case studies in the application of principal component analysis. Applied Statistics, 16(3) :225–236.

John, G., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. Proceedings of the Eleventh International Conference on Machine Learning, 129.

Jolliffe, I. (2002). Principal component analysis. 2nd edition, Springer-Verlag New York.

Jolliffe, I., Trendafilov, N., & Uddin, M. (2003). A Modified Principal Component Technique Based on the LASSO. Journal of Computational & Graphical Statistics, 12(3) :531–547.

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., & Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med, 7(6) :673–679.

Kim, H., Pang, S., Je, H., Kim, D., & Bang, S. (2002). Pattern Classification Using Support Vector Machine Ensemble. Proc. of ICPR, 2 :20160–20163.

Kira, K. & Rendell, L. (1992). The Feature Selection Problem : Traditional Methods and a New Algorithm. AAAI-92 : Proceedings Tenth National Conference on Artificial Intelligence/July 12-16, 1992.

Kohavi, R. & John, G. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97(1-2) :273–324.

Kohonen, T. (2001). Self-Organizing Maps. Springer.

Kononenko, I. (1994). Estimating attributes : Analysis and extensions of RELIEF. Proceedings of the European Conference on Machine Learning, pages 171–182.

Kuss, M. & Graepel, T. (2003). The geometry of canonical correlation analysis. Technical report, Max Planck Institute for Biological Cybernetics.

Lê Cao, K.-A. (2005). Discrimination et sélection de variables appliquées à des données de biopuces : application à la folliculogénèse chez le porc. Master's thesis, Université de Toulouse et Institut National de la Recherche Agronomique.

Lê Cao, K.-A., Bonnet, A., & Gadat, S. (2007a). Multiclass classification and gene selection with a stochastic algorithm. Technical report, Université de Toulouse, Institut National de la Recherche Agronomique.

Lê Cao, K.-A., Gonçalves, O., Besse, P., & Gadat, S. (2007b). Selection of biologically relevant genes with a wrapper stochastic algorithm. Statistical Applications in Genetics and Molecular Biology, 6(:Iss. 1) :Article 1.

Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., & Besse, P. (2008). Sparse PLS : Variable Selection when Integrating Omics data. Technical report, Université de Toulouse et Institut National de la Recherche Agronomique.

Lee, J., Lee, J., Park, M., & Song, S. (2005). An extensive comparison of recent classification tools applied to microarray data. Computational Statistics and Data Analysis, 48(4) :869–885.

Lee, M., Whitmore, G., & Yukhananov, R. (2003). Analysis of Unbalanced Microarray Data. Journal of Data Science, 1 :103–121.

Lee, Y. & Lee, C. (2003). Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics, 19(9) :1132–1139.

Leurgans, S. E., Moyeed, R. A., & Silverman, B. W. (1993). Canonical correlation analysis when the data are curves. Journal of the Royal Statistical Society B, 55(3) :725–740.

Li, T., Zhang, C., & Ogihara, M. (2004a). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20(15) :2429–2437.

Li, T., Zhang, C., & Ogihara, M. (2004b). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20(15) :2429–2437.

Liaw, A. & Wiener, M. (2002). Classification and regression by randomforest. Rnews, 2/3(December) :18–22.

Lorber, A., Wangen, L., & Kowalski, B. (1987). A theoretical foundation for the PLS algorithm. Journal of Chemometrics, 1(19-31) :13.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). Multivariate Analysis. Academic Press.

Mevik, B.-H. & Wehrens, R. (2007). The pls package : Principal component and partial least squares regression in r. Journal of Statistical Software, 18(2).

Mundra, P. & Rajapakse, J. (2007). SVM-RFE with Relevancy and Redundancy Criteria for Gene Selection. Lecture notes in computer science, 4774 :242.

Nguyen, D. & Rocke, D. (2004). On partial least squares dimension reduction for microarray-based classification : a simulation study. Computational Statistics and Data Analysis, 46(3) :407–425.

Nykter, M., Aho, T., Ahdesmaki, M., Ruusuvuori, P., Lehmussola, A., & Yli-Harja, O. (2006). Simulation of microarray data with realistic characteristics. BMC Bioinformatics, 7(1) :349.

Pirooznia, M., Yang, J., Yang, M. Q., & Deng, Y. (2008). A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics, 9(Suppl 1) :S13.

Prasad, A., Iverson, L., & Liaw, A. (2006). Newer Classification and Regression Tree Techniques : Bagging and Random Forests for Ecological Prediction. Ecosystems, 9(2) :181–199.

Qiao, X. & Liu, Y. (2008). Adaptive Weighted Learning for Unbalanced Multicategory Classification. Biometrics.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences, 98(26) :15149–15154.

Reunanen, J., Guyon, I., & Elisseeff, A. (2003). Overfitting in Making Comparisons Between Variable Selection Methods. Journal of Machine Learning Research, 3(7-8) :1371–1382.

Robnik-Sikonja, M. & Kononenko, I. (1997). An adaptation of Relief for attribute estimation in regression. Machine Learning : Proceedings of the Fourteenth International Conference (ICML 97), pages 296–304.

Sampson, P., Streissguth, A., Barr, H., & Bookstein, F. (1989). Neurobehavioral effects of prenatal alcohol : Part II. Partial Least Squares analysis. Neurotoxicology and Teratology, 11(5) :477–491.

Saporta, G. (2006). Probabilités analyses des données et Statistiques. Editions Technip, Paris.

Scholkopf, B. & Smola, A. (2001). Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press Cambridge, MA, USA.

Shen, H. & Huang, J. Z. (2007). Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis, to appear.

Simon, R., Radmacher, M., Dobbin, K., & McShane, L. (2003). Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. jnci, 95(1) :14–18.

Sorensen, P., Bonnet, A., Buitenhuis, B., Closset, R., Dejean, S., Delmas, C., Duval, M., Glass, L., Hedegaard, J., Hornshoj, H., et al. (2007). Analysis of the real EADGENE data set : multivariate approaches and post analysis (open access publication). Genet Sel Evol, 39(6) :651–68.

Statnikov, A., Aliferis, C., Tsamardinos, I., Hardin, D., & Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics, 21(5) :631–643.

Steinfath, M., Groth, D., Lisec, J., & Selbig, J. (2008). Metabolite profile analysis : from raw data to regression and classification. Physiologia Plantarum, 132(2) :150–161.

Streissguth, A., Bookstein, F., Sampson, P., & Barr, H. (1993). The Enduring Effects of Prenatal Alcohol Exposure on Child Development, Birth Through 7 Years : A Partial Least Squares Solution. Ann Arbor : University of Michigan Press, 301.

Strobl, C., Boulesteix, A., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures : Illustrations, sources and a solution. BMC Bioinformatics, 8(1) :25.

Su, Y., Murali, T., Pavlovic, V., Schaffer, M., & Kasif, S. (2003). RankGene : identification of diagnostic genes based on expression data. Bioinformatics, 19(12) :1578–1579.

Sun, Y. & Li, J. (2006). Iterative RELIEF for feature weighting. Proceedings of the 23rd international conference on Machine learning, pages 913–920.

Svetnik, V., Liaw, A., & Tong, C. (2003). Variable Selection in Random Forest with Application to Quantitative Structure-Activity Relationship. Technical report, Working paper, Biometrics Research Group, Merck & Co., Inc.

Tang, Y., Zhang, Y.-Q., & Huang, Z. (2007). Development of two-stage svm-rfe gene selection strategy for microarray expression data analysis. IEEE/ACM Trans. Comput. Biol. Bioinformatics, 4(3) :365–381.

Tenenhaus, M. (1998). La régression PLS : théorie et pratique. Editions Technip.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 58(1) :267–288.

Tosser-Klopp, G., Lê Cao, K.-A., Bonnet, A., Gobert, N., Hatey, F., Robert-Granié, C., Déjean, S., Antic, J., Baschet, L., & San Cristobal, M. (2008). A pilot study on transcriptome data analysis of folliculogenesis in pigs. Reproduction.

Trendafilov, N. & Jolliffe, I. (2006). Projected gradient approach to the numerical solution of the SCoTLASS. Computational Statistics and Data Analysis, 50(1) :242–253.

Trendafilov, N. & Jolliffe, I. (2007). DALASS : Variable selection in discriminant analysis via the LASSO. Computational Statistics and Data Analysis, 51(8) :3718–3736.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6) :520–525.

Trygg, J. & Wold, S. (2003). O2-pls, a two- block (x-y) latent variable regression (lvr) method with an integral osc filter. Journal of Chemometrics, 17 :53–64.

Valafar, F. (2002). Pattern Recognition Techniques in Microarray Data Analysis A Survey. Annals of the New York Academy of Sciences, 980(1) :41–64.

Vapnik, V. N. (1999). The Nature of Statistical Learning Theory (Information Science and Statistics). Springer.

Vert, J. & Kanehisa, M. (2003). Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA. Advances in Neural Information Processing Systems, 15 :1425–1432.

Vinod, H. D. (1976). Canonical ridge and econometrics of joint production. Journal of Econometrics, 4(2) :147–166.

Waaijenborg, S., de Witt Hamer, V., Philip, C., & Zwinderman, A. (2008). Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis. Statistical Applications in Genetics and Molecular Biology, 7(1) :3.

Watson, M., Pérez-Alegre, M., Baron, M., Delmas, C., Dovc, P., Duval, M., Foulley, J., Garrido-Pavón, J., Hulsegge, I., Jaffrézic, F., et al. (2007). Analysis of a simulated microarray dataset : Comparison of methods for data normalisation and detection of differential expression (Open Access publication). Genet. Sel. Evol, 39 :669–683.

Wegelin, J. (2000). A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Technical Report 371, Department of Statistics, University of Washington, Seattle.

Weston, J., Elisseeff, A., Schölkopf, B., & Tipping, M. (2003). Use of the zero norm with linear models and kernel methods. J. Mach. Learn. Res., 3 :1439–1461.

Weston, J. & Watkins, C. (1999). Support vector machines for multi-class pattern recognition. Proceedings of the Seventh European Symposium On Artificial Neural Networks, 4 :6.

Wiesel, A., Kliger, M., & Hero III, A. (2008). A greedy approach to sparse canonical correlation analysis. Arxiv preprint arXiv :0801.2748.

Wold, H. (1966). Multivariate Analysis. Academic Press, New York, Wiley, krishnaiah, p.r. (ed.) edition.

Wold, S., Eriksson, L., Trygg, J., & Kettaneh, N. (2004). The PLS method–partial least squares projections to latent structures–and its applications in industrial RDP (research, development, and production). Technical report, Umea University.

Yamanishi, Y., Vert, J., Nakaya, A., & Kanehisa, M. (2003). Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. Bioinformatics, 19(90001) :323–330.

Yeang, C., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R., Angelo, M., Reich, M., Lander, E., Mesirov, J., & Golub, T. (2001). Molecular classification of multiple tumor types. Bioinformatics, 17(90001) :316–322.

Yeung, K. & Burmgarner, R. (2003). Multi-class classification of microarray data with repeated measurements : application to cancer. Genome Biology, 4(R83).

Yousef, M., Jung, S., Showe, L., & Showe, M. (2007). Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data. BMC Bioinformatics, 8(1) :144.

Zeng, X.-Q., Li, G.-Z., Yang, J., Yang, M., & Wu, G.-F. (2008). Dimension reduction with redundant gene elimination for tumor classification. BMC Bioinformatics, 9(Suppl 6) :S8.

Zhang, H., Liu, Y., Wu, Y., & Zhu, J. (2008). Variable selection for multicategory SVM via sup-norm regularization. Electronic Journal of Statistics, 2 :149–167.

Zhou, X. & Tuck, D. (2007). MSVM-RFE. Bioinformatics, 23(9) :1106–1114.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B, 67(2) :301–320.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. Journal of Computational and Graphical Statistics, 15(2) :265–286.