Sélection de variables pour la classification par mélanges gaussiens pour prédire la fonction des gènes orphelins

Cathy Maugis[1], Marie-Laure Martin-Magniette[2,3], Jean-Philippe Tamby[3], Jean-Pierre Renou[3], Alain Lecharny[3], Sébastien Aubourg[3], Gilles Celeux[4]

[1] Département de Mathématiques, Université Paris-Sud 11, Orsay, France
E-mail: cathy.maugis@math.u-psud.fr
[2] UMR AgroParisTech/INRA MIA 518, Paris, France
E-mail: marie_laure.martin@agroparistech.fr
[3] URGV UMR INRA 1165, CNRS 8114, UEVE, Evry, France
E-mail: aubourg@evry.inra.fr; lecharny@evry.inra.fr; renou@evry.inra.fr; tamby@versailles.inra.fr
[4] INRIA Saclay Île-de-France, France
E-mail: gilles.celeux@inria.fr

Abstract Biologists are interested in predicting the gene functions of sequenced genome organisms according to microarray transcriptome data. The microarray technology development allows one to study the whole genome in different experimental conditions. The information abundance may seem to be an advantage for the gene clustering. However, the structure of interest can often be contained in a subset of the available variables. The currently available variable selection procedures in model-based clustering assume that the irrelevant clustering variables are all independent or are all linked with the relevant clustering variables. A more versatile variable selection model is proposed, taking into account three possible roles for each variable: The relevant clustering variables, the redundant variables and the independent variables. A model selection criterion and a variable selection algorithm are derived for this new variable role modelling. The interest of this new modelling for discovering the function of orphan genes is highlighted on a transcriptome dataset for the *Arabidopsis thaliana* plant.
Keywords: Variable selection, model-based clustering, transcriptome data, orphan genes.

Résumé Les biologistes s'attachent actuellement à prédire la fonction des gènes d'organismes de génome séquencé à partir de données transcriptomes, issues de l'utilisation des puces à ADN. Le développement de cette technologie permet de tester l'expression de l'ensemble du génome dans de nombreuses conditions expérimentales. Cette quantité d'information peut alors sembler être un atout pour la classification des gènes. Pourtant il est courant que seul un sous-ensemble contienne l'information pertinente pour la classification. Les procédures de sélection des variables en classification non supervisée par mélanges gaussiens supposent généralement que les variables non informatives pour la classification sont soit toutes indépendantes, soit liées à des variables informatives. Nous proposons une nouvelle modélisation du rôle des variables plus polyvalente : les variables sont soit informatives pour la classification, soit redondantes, soit totalement indépendantes. Nous proposons un critère de sélection des variables et un algorithme pour cette nouvelle modélisation. L'intérêt

de cette nouvelle modélisation pour la prédiction de la fonction des gènes orphelins est illustrée sur un ensemble de données transcriptomes obtenues chez *Arabidopsis thaliana*.
Mots clés : Sélection de variables, mélanges gaussiens, données transcriptomes, gènes orphelins.

1 Introduction

Malgré l'augmentation du nombre de génomes complètement séquencés et les progrès techniques dans les expériences de génomique, l'annotation fonctionnelle qui consiste à déterminer la fonction des gènes reste difficile pour 15 à 40% des gènes (Gollery et al., 2006). Jusqu'à présent, cette limite est vraie pour tous les génomes, des bactéries aux mammifères. Parmi les six plantes dont le génome est publié (*Arabidopsis thaliana*, le riz, la mousse *Physcomitrella patens*, le peuplier, l'algue *Ostreococcus tauri* et la vigne), l'annotation du génome d'*Arabidopsis* a l'avantage de bénéficier d'efforts importants depuis sa première version il y a huit ans (AGI, 2000). Néanmoins, après sept versions officielles de l'annotation, environ 4000 gènes d'*Arabidopsis* sont toujours considérés orphelins, c'est-à-dire sans fonction biologique connue ou prédite.

Depuis une dizaine d'années, la technologie des puces à ADN permet de générer un nombre considérable de données sur les transcrits d'un organisme. Le principe est d'avoir sur un support miniaturisé tous les gènes d'un organisme et d'hybrider sur ce support deux échantillons biologiques (les ARN messagers) obtenus dans des conditions expérimentales différentes. L'analyse différentielle de ces données permet d'identifier les gènes qui s'expriment de manière différente entre les deux conditions. Avec la vaste utilisation de cette technologie, les données du transcriptome constituent la principale source d'information et il est désormais possible de concevoir une approche globale pour déterminer la fonction de ces gènes orphelins. L'hypothèse, proposée pour la première fois par Eisen et al. (1998), que des gènes partageant un même profil d'expression sont très certainement impliqués dans un même processus biologique est généralement admise. Le principe consiste alors à utiliser des méthodes de classification non supervisée sur un ensemble de gènes, composé des gènes orphelins et de gènes connus pour être impliqués dans un processus d'intérêt, afin de constituer des groupes de gènes ayant le même profil d'expression sur un ensemble d'expériences transcriptomes. Ainsi la mise en évidence de groupes de gènes co-exprimés permet d'aider les biologistes et les bioinformaticiens à identifier des gènes co-régulés et à proposer une fonction biologique aux gènes orphelins, en caractérisant biologiquement les classes identifiées par la classification non supervisée.

Pour déterminer des classes d'observations décrites par des variables quantitatives, nous considérons la classification par mélanges gaussiens. On constate un regain de popularité à l'égard de cette méthode de classification, la flexibilité des mélanges gaussiens permettant de modéliser une large variété de phénomènes aléatoires. Cette attention est due au fait que ces mélanges reflètent l'idée intuitive que l'ensemble des observations est composée de plusieurs sous-populations, chacune modélisée par une densité gaussienne multidimensionnelle. L'avantage de cette méthode est de fournir un cadre statistique rigoureux pour déterminer le nombre de composants du mélange et décrire le rôle des variables dans le processus de classification. Mais cette méthode peut être mise en difficulté lors de l'étude de données de grande dimension. En effet, il est généralement admis que plus le nombre de variables pour décrire les observations est important, plus il est

aisé de faire de la classification non supervisée. Or quand le nombre de variables est très important, la structure d'intérêt pour la classification peut être contenue que dans un sous-ensemble de variables, les autres variables étant inutiles, redondantes voire néfastes pour détecter une classification convenable des données. Il est donc important de déterminer le rôle des variables vis-à-vis de la classification et de détecter en particulier les variables informatives. Ce problème de sélection de variables en classification non supervisée est un sujet récent motivé par l'utilisation de plus en plus intensive de méthodes de classification pour l'étude de données de grande dimension telles que les données transcriptomes. Une revue des trois types d'approches existantes est proposée dans l'article de Maugis et al. (2009a). Dans le cadre de l'étude de la fonction des gènes orphelins, le nombre de données transcriptomes augmentant sans cesse, la sélection des expériences informatives pour la classification est souhaitable pour regrouper les gènes qui collaborent dans un même processus biologique.

Dans les différents articles abordant la sélection de variables dans le cadre de la classification non supervisée par mélanges gaussiens, le problème est reformulé en un problème de sélection de modèles. Parmi les procédures déjà proposées, celle de Law et al. (2004) est fondée sur l'hypothèse que les variables non informatives pour la classification sont totalement indépendantes des variables significatives. Pour remédier cette hypothèse restrictive, Raftery et Dean (2006) proposent une modélisation où les variables non informatives sont supposées liées à toutes les variables informatives selon une régression linéaire. Une généralisation de la modélisation de Raftery et Dean est proposée par Maugis et al. (2009a) en permettant que les variables non informatives soient expliquées par un sous-ensemble des variables informatives selon une régression linéaire. Cette nouvelle modélisation englobe les cas particuliers envisagés par Raftery et Dean (2006) et Law et al. (2004). Cependant cette modélisation du rôle des variables n'est pas complètement générale puisqu'elle ne permet pas d'avoir simultanément des variables non informatives indépendantes et d'autres dépendantes de variables significatives. Nous avons ainsi proposé dans Maugis et al. (2009b) d'affiner la modélisation pour permettre à chaque variable d'être soit informative pour la classification, soit redondante ou soit indépendante. Cette nouvelle modélisation permet d'améliorer la classification des données et de fournir simultanément une analyse du rôle des variables (voir Maugis et al., 2009b). Dans cet article, nous reprenons cette modélisation et présentons son application à la prédiction de la fonction de gènes orphelins.

2 Modélisation du rôle des variables

Nous considérons un n-échantillon de loi inconnue $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ décrit par Q variables quantitatives. Notre objectif consiste à combiner une procédure de sélection de variables au processus de classification par mélanges gaussiens afin de déterminer des classes d'observations et faciliter leur interprétation. Ce problème est abordé comme un problème de sélection de modèles, la modélisation étant axée sur le rôle des variables vis-à-vis de la classification. Nous commençons par noter S l'ensemble non vide des variables informatives pour la classification et S^c son complémentaire, contenant les variables non informatives. Ce dernier est partitionné en deux sous-ensembles U et W. Les variables appartenant à U, dites redondantes, sont expliquées par un sous-ensemble R de S à l'aide d'une régression linéaire tandis que les variables dans W, dites indépendantes, sont

supposées indépendantes de toutes les variables informatives pour la classification.

Pour une partition de variables (S, R, U, W), la densité inconnue de l'échantillon est alors modélisée par une densité définie comme le produit des trois termes suivants :

- sur l'ensemble S des variables informatives pour la classification, nous considérons un mélange gaussien caractérisé par son nombre de composants K et sa forme m principalement fondée sur des contraintes imposées sur les termes de la décomposition spectrale des matrices de variance (Banfield and Raftery, 1993; Celeux and Govaert, 1995).
- Les variables de l'ensemble U sont expliquées par les variables de l'ensemble R à l'aide d'une régression linéaire multidimensionnelle. La forme de la matrice de variance de cette dernière, notée r, est soit sphérique, soit diagonale, soit générale.
- La distribution marginale des données décrites par les variables indépendantes du sous-ensemble W est supposée gaussienne. La forme de la matrice de variance de la densité gaussienne, notée l, est soit sphérique, soit diagonale.

La collection de modèles considérée est donc composée de modèles décrits par le uplet (K, m, r, l, S, R, U, W). Cette modélisation, désignée SRUW, permet de considérer une collection plus riche en information que la modélisation SR proposée par Maugis et al. (2009a). En effet, tout modèle de la modélisation SR peut être vu comme un modèle de SRUW avec $U = S^c$ et $W = \emptyset$. L'identifiabilité de cette nouvelle collection de modèles est établie et énoncée dans Maugis et al. (2009b). Les conditions requises sont similaires à celles utilisées par Maugis et al. (2009a) pour l'identifiabilité des modèles de la modélisation SR.

3 Critère de sélection de modèles

La collection de modèles SRUW permet de concevoir le problème de sélection de variables pour la classification par mélanges gaussiens en un problème de sélection de modèles. Idéalement les modèles en compétition sont comparés à l'aide de leur log-vraisemblance intégrée. En pratique, ces log-vraisemblances intégrées étant difficilement calculables elles sont approchées par le critère BIC (Bayesian Information Criterion, Schwarz, 1978) et le meilleur modèle de la collection est celui qui maximise le critère BIC

$$2 \times \texttt{maximum de log-vraisemblance} - [\texttt{degrés de liberté} \times \ln(n)] .$$

Ce critère de sélection de modèle BIC peut être décomposé comme la somme de trois critères BIC associés respectivement au mélange gaussien, à la régression linéaire et à la distribution gaussienne des données décrites par les variables indépendantes. Les détails sur l'expression de ce critère et l'estimation des paramètres sont décrits dans Maugis et al. (2009b). La consistance de ce critère de sélection de variables est également établie dans Maugis et al. (2008) sous des conditions classiques de régularité.

4 La procédure de sélection de variables

La taille de la collection de modèles considérée est telle qu'une recherche exhaustive du meilleur modèle est impossible. Nous avons donc proposé un algorithme, appelé *Sel-*

varClustIndep (disponible sur `http://www.math.u-psud.fr/~maugis`), fondé sur deux algorithmes *backward-stepwise* emboîtés pour déterminer la meilleure partition des variables. Lors d'une étape de l'algorithme de sélection des variables, trois situations peuvent être envisagées pour une variable susceptible d'être incluse ou exclue de l'ensemble des variables informatives pour la classification. L'astuce consiste à ramener la comparaison des trois situations en une comparaison de deux situations pour la modélisation SR (voir Maugis et al., 2009b). Ainsi pour chaque modèle de mélange défini par le couple (K, m), la première étape de l'algorithme consiste à séparer les variables informatives pour la classification des variables non informatives à l'aide de l'algorithme *SelvarClust* associé à la modélisation SR (disponible sur `http://www.math.u-psud.fr/~maugis`). On obtient ainsi les ensembles $S(K, m)$ et $S^c(K, m)$. Dans la seconde étape, les variables non informatives sont réparties dans les ensembles $U(K, m)$ et $W(K, m)$: pour chaque variable de $S^c(K, m)$, le sous-ensemble des variables significatives nécessaires pour l'expliquer est déterminé par un algorithme backward stepwise de régression et la variable testée est déclarée redondante si le sous-ensemble est non vide et indépendante sinon. Ensuite, l'ensemble des variables informatives $R(K, m, r)$ pour la régression multidimensionnelle des variables redondantes et les formes des matrices de variance du modèle de régression et de la gaussienne modélisant les données décrites par les variables indépendantes sont déterminés. La dernière étape consiste alors à sélectionner le modèle (K, m, r, l) maximisant notre critère avec la partition de variables associée $(S(K, m), R(K, m, l), U(K, m), W(K, m))$.

Il convient de noter que la complexité de l'algorithme *SelvarClustIndep* est du même ordre que celle de l'algorithme associé à la modélisation SR, malgré les trois rôles de variables. L'utilisation de cet algorithme dans Maugis et al. (2009b) pour l'étude d'un jeu de données simulées avec différents scénarios pour les variables non informatives et les données *waveforms* (Breiman et al., 1984) a permis d'illustrer le comportement de la modélisation SRUW et de la comparer à la méthode de sélection SR. Nous avons pu montrer que la modélisation SRUW permet de construire des classes d'individus plus homogènes et que la partition des variables en variables informatives, redondantes et indépendantes apportent des informations utiles pour l'interprétation des résultats.

5 Analyse d'un jeu de données transcriptomes

L'objectif de cette section est d'illustrer l'utilisation de la modélisation SRUW dans le cadre de l'étude de données transcriptomes. Dans cet article, nous considérons des données transcriptomes pour la plante modèle *Arabidopsis thaliana* extraites de la base de données CATdb développée par Gagnot et al. (2008) à l'Unité de Recherche en Génomique Végétale (URGV). L'avantage de ces données est qu'elles sont toutes produites avec la puce à ADN CATMA (Crowe et al., 2003) sur la même plateforme transcriptome et avec le même protocole. De plus les analyses statistiques intégrées dans CATdb sont menées de façon identiques pour chaque expérience. Elles consistent à éliminer les biais techniques (normalisation) et à déterminer les gènes significativement différentiellement exprimés (analyse différentielle) entre deux conditions. Le lecteur intéressé peut se référer à l'article de Gagnot et al. (2008) pour une description détaillée de ces analyses statistiques. Ainsi d'un point de vue statistique, la variabilité due aux effets techniques est contrôlée et peut donc être considérée comme homogène pour toutes les expériences.

Le jeu de données étudié consiste en 4616 gènes d'*Arabidopsis thaliana*, dont 1430 gènes orphelins (de fonction inconnue), décrits par $Q = 33$ expériences de stress biotiques regroupées en neuf projets. Chaque projet est composé d'un ensemble d'expériences dédiées à une question biologique spécifique. Chaque gène est décrit par un vecteur $\mathbf{y}_i \in \mathbb{R}^Q$ dont la $j$ième composante correspond à la valeur de la statistique de test du gène i dans l'expérience j lors de l'analyse différentielle. Cette statistique de test est définie comme une différence d'expression normalisée. Nous avons mené l'étude avec un nombre K de composants pour le mélange variant de 10 à 30 et les formes de mélange $[p_k LC]$ (proportions différentes et matrices variance générales identiques) et $[p_k L_k C]$ (proportions différentes et pour les matrice variance, volumes différents, orientations et formes identiques). Ce choix restreint seulement à deux formes est motivé par des études antérieures sur des données transcriptomes dont celle proposée dans Maugis et al. (2009a). Il faut noter que l'estimation des paramètres du mélange gaussien se fait à plusieurs reprises dans chaque étape de l'algorithme. Aussi, il arrive parfois que la procédure n'aboutisse pas pour une forme particulière. Par exemple, la procédure de sélection de variables a ici parfois échoué pour des valeurs de K sous la forme $[p_k L_k C]$ bien que les meilleures résultats sont obtenus sous cette forme d'après les valeurs du critère de sélection pour l'étude du jeu de données transcriptomes considéré. Notre algorithme *SelvarClustIndep* a finalement sélectionné un mélange gaussien de forme $m = [p_k L_k C]$ avec $\hat{K} = 26$ classes, certaines étant communes à la classification en 23 classes obtenue sans sélection de variables. Les expériences 22, 23 et 26 sont déclarées redondantes et toutes les autres significatives pour la classification. Parmi les 26 classes de tailles différentes (voir Tab. 2), certains sous-groupes de gènes coexprimés sont mis en évidence. Par exemple, les profils d'expression des gènes de la classe 19, représentés sur la figure Fig. 1, sont très homogènes. Cette cohérence d'expression des gènes est particulièrement visible sur la représentation des log-ratios en vert, rouge et noire utilisée par les biologistes (voir Fig. 1 à droite). Ainsi on peut par exemple voir que ces gènes, qui ont des statistiques de test fortement positives dans l'expérience 27, sont déclarés avoir une différence d'expression positive (rouge) dans le tableau des log-ratios.

Grâce à la classification par mélanges gaussiens, on peut mesurer l'adéquation d'appartenance du gène i à une classe G_k par les probabilités conditionnelles d'appartenance $P(i \in G_k|\mathbf{y})$. Le tableau Tab. 1 représente la répartition des gènes selon la valeur de

$$\text{pmax}(i) := \max_{1 \leq k \leq K} P(i \in G_k|\mathbf{y}).$$

On peut alors choisir un seuil c pour distinguer les gènes dits « sûrs » ($\text{pmax}(i) \geq c$) des gènes dits « mitigés » ($\text{pmax}(i) < c$). Cette quantification par les probabilités conditionnelles d'appartenance permet ainsi aux biologistes de mesurer l'homogénéité et la pertinence d'une classe. Dans le tableau Tab. 2, la distinction des gènes en « sûrs » et « mitigés » pour chaque classe est donnée pour le seuil $c = 0.8$.

pmax	$[0.2, 0.3[$	$[0.3, 0.4[$	$[0.4, 0.5[$	$[0.5, 0.6[$	$[0.6, 0.7[$	$[0.7, 0.8[$	$[0.8, 0.9[$	$[0.9, 1]$
Nombre de gènes	1	67	189	321	358	453	638	2589

Tab. 1 – Répartition des gènes selon la valeur de pmax.

Numéro de classe	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	Total
Nombre de gènes	133	113	277	87	78	303	67	221	151	79	329	255	602	214	80	127	45	60	129	165	333	123	156	109	264	116	4616
Nombre de gènes « sûrs »	112	79	221	60	60	200	50	140	103	57	219	187	313	158	72	105	37	42	108	120	238	88	98	97	178	85	3227
Nombre de gènes « mitigés »	21	34	56	27	18	103	17	81	48	22	110	68	289	56	8	22	8	18	21	45	95	35	58	12	86	31	1389
Nombre de gènes orphelins	34	35	78	24	26	94	14	68	41	25	129	103	192	60	27	35	14	16	8	50	124	26	56	36	76	39	1430

TAB. 2 – Répartition des gènes selon les 26 classes.

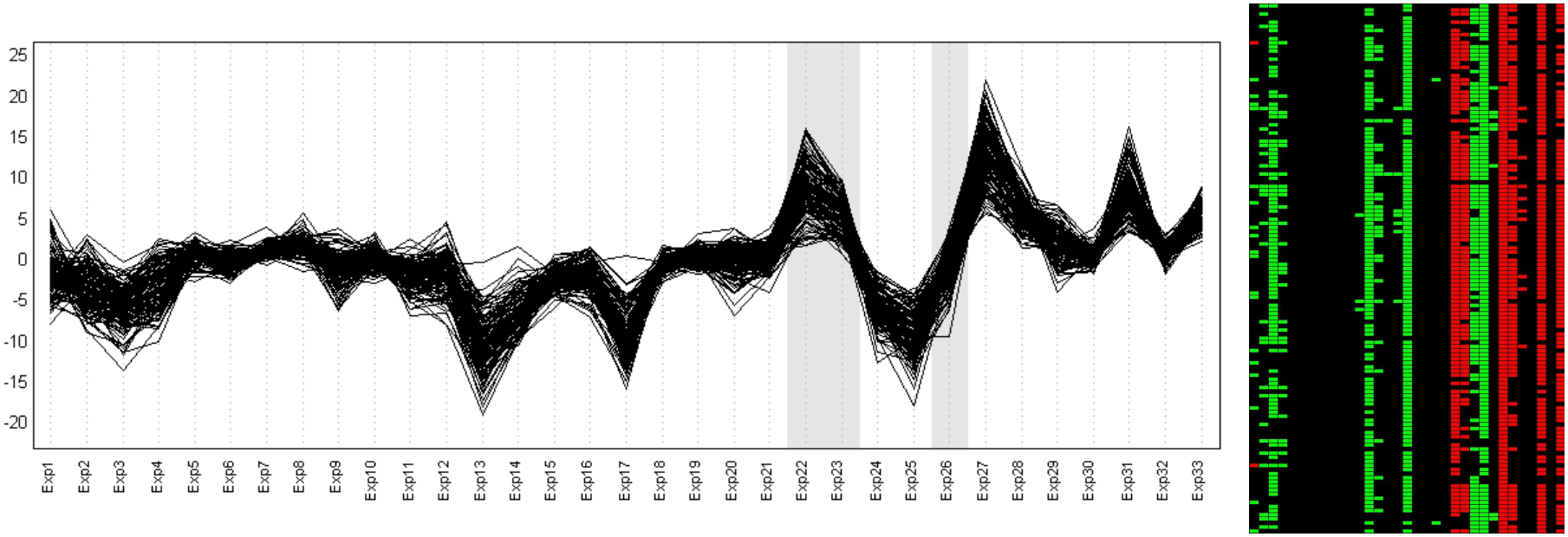

FIG. 1 – À gauche, représentation des profils d'expression des gènes de la classe 19 (les variables redondantes sont représentées sur fond gris). À droite, la représentation des log-ratios utilisée couramment par les biologistes : chaque ligne correspond à un gène et chaque colonne à une expérience. La case est rouge si le gène a un log-ratio positif dans l'expérience concernée (différence d'expression positive, « Up »), verte si le log-ratio est négatif (différence d'expression négative, « Down ») et noire sinon.

Concernant la sélection de variables, on peut remarquer que les trois expériences 22, 23 et 26 déclarées redondantes sont issues d'un même projet biologique. Ces expériences sont régressées selon les variables $\hat{R} = \{1, 5, 8, 13, 24, 25, 27, 28-31, 33\}$. D'après les paramètres estimés de la régression linéaire (voir TAB. 3), ces trois expériences sont très corrélées et sont principalement expliquées par les expériences 24, 25 et 27 qui font partie du même projet biologique.

	Exp 22	Exp 23	Exp 26
constante	-0.050997	-0.066629	-0.187973
Exp 1	0.001552	-0.01585	-0.018604
Exp 5	0.040455	0.015224	0.046560
Exp 8	-0.051241	-0.029414	-0.073089
Exp 13	-0.005727	-0.019719	-0.031408
Exp 24	0.916864	0.007923	1.006624
Exp 25	-0.028760	0.806759	0.836472
Exp 27	1.006091	1.000483	0.998316
Exp 28	-0.022752	-0.030231	-0.057648
Exp 29	0.028125	0.047300	0.056751
Exp 30	0.056352	0.070486	0.094796
Exp 31	0.025773	-0.034398	-0.006417
Exp 33	0.044237	-0.010248	-0.016750

	Exp 22	Exp 23	Exp 26
Exp 22	1.245884	0.458902	0.960841
Exp 23	0.458902	1.209113	0.978319
Exp 26	0.960841	0.978319	2.060182

	Exp 22	Exp 23	Exp 26
Exp 22	1.0000	0.4173	0.6694
Exp 23	0.4173	1.0000	0.6816
Exp 26	0.6694	0.6816	1.000

TAB. 3 – Paramètres estimés de la régression : à gauche, la matrice des coefficients de régression et à droite, la matrice de covariance (en haut) et la matrice de corrélation (en bas).

Pour la validation biologique de ces résultats, nous avons recours à des données issues de la base de données FLAGdb^{++} (Samson et al., 2004). L'annotation fonctionnelle des protéines codées par les gènes d'une même classe est un premier élément pour juger de la pertinence de la classification. Cependant, cette annotation fonctionnelle est dépendante de la présence de protéines homologues de fonction connue, elle concerne essentiellement la fonction biochimique des protéines et rarement la fonction physiologique, plus étroitement reliée à la transcription des gènes. Pour évaluer la cohérence biologique des classes, nous avons donc étudié la localisation subcellulaire prédite des protéines codées par les gènes de chaque classe. En effet, des protéines qui collaborent fonctionnellement auront tendance à interagir au sein d'un même compartiment cellulaire. S'appuyant sur la détection bioinformatique des signaux d'adressage dont la structure et la composition sont bien connues, la figure FIG. 2 présente la distribution des protéines dans le noyau, le réticulum endoplasmique, les plastes et les mitochondries. De nombreuses classes se différencient de la tendance générale (« Référence » sur le graphique) et affichent des biais de localisation très marqués dans les quatre compartiments : plastes (classes 12, 20 et 23), réticulum endoplasmique (classes 7 et 16), noyau (classe 19) et mitochondries (classe 3). Ces résultats constituent un bon indicateur de la robustesse biologique des classes obtenues.

Si on s'intéresse plus particulièrement aux gènes de la classe 19, on remarque que 101 de ces gènes ont une fonction liée au phénomène de traduction et 8 sont des gènes orphelins d'après l'annotation fonctionnelle. D'après la figure FIG. 2, une grande partie des protéines codées par ces gènes ont un adressage nucléaire. Ceci permet aux constituants protéiques des ribosomes d'accéder au nucléole (sous-compartiment du noyau) dans lequel ils vont s'associer aux ARN ribosomiques. Les ribonucléoprotéines obtenues ressortent ensuite du noyau pour s'assembler et former les ribosomes qui permettent la traduction. Ces

premiers résultats permettent alors aux biologistes d'émettre de nouvelles hypothèses sur la fonction des 8 gènes orphelins qui devront par la suite être validées expérimentalement.

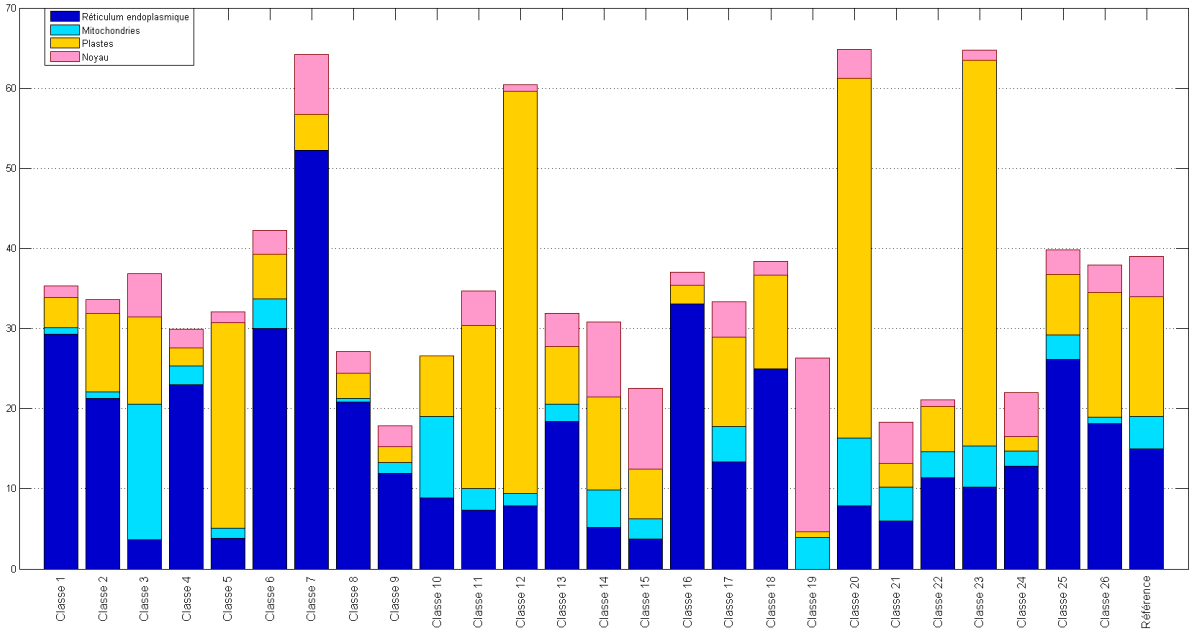

FIG. 2 – Répartition dans le réticulum endoplasmique, les mitochondries, les plastes et le noyau des protéines codées par les gènes de chaque classe. La colonne « Référence » représente la répartition de toutes les protéines d'*Arabidopsis thaliana*.

Parmi les 26 classes obtenues pour la classification des 4616 gènes, certaines restent inexploitables par les biologistes car elles rassemblent un nombre élevé de gènes dont les profils d'expression sont moins homogènes. On peut alors envisager d'étudier ces classes séparément avec notre procédure de sélection de variables afin d'en extraire de l'information. Nous nous sommes ici intéressés à l'étude de la classe 11 composée de 329 gènes. L'algorithme *SelvarClustIndep* a été appliqué avec K variant de 2 à 10 et les formes $[p_k LC]$ et $[p_k L_k C]$. Les 329 gènes ont été répartis en 5 classes et la forme sélectionnée pour le mélange gaussien est $[p_k LC]$. Les expériences 1, 5 et 20 sont déclarées significatives pour la classification, les expériences 9 et 18 sont indépendantes et les autres sont redondantes, régressées selon l'expérience 1. Cette classification permet de mettre en évidence de petites classes, dites sous-classes 1, 2, 4 et 5 composées de 36, 24, 11 et 10 gènes respectivement. On peut tout d'abord noter que puisque les gènes de la classe initiale 11 ont un comportement caractéristique pour les expériences 25 à 33 (voir FIG. 4), ces expériences ne sont pas déclarées significatives car elles ne permettent plus de distinguer ces 329 gènes. On peut également remarquer que les gènes ont été classés selon leur différence d'expression dans l'expérience 1 : les gènes ont une différence d'expression positive dans les sous-classes 2, 3 et 5 et négatives dans les sous-classes 1 et 4 (voir FIG. 3). On observe aussi que les gènes des sous-classes 2 et 4 ont des profils caractéristiques intéressants dans l'expérience 20. Enfin, les expériences 9 et 18 déclarées indépendantes sont les expériences qui sont les moins corrélées aux expériences significatives. Le nombre de variables indépendantes reste faible, soulignant ainsi les relations complexes entre les données transcriptomes.

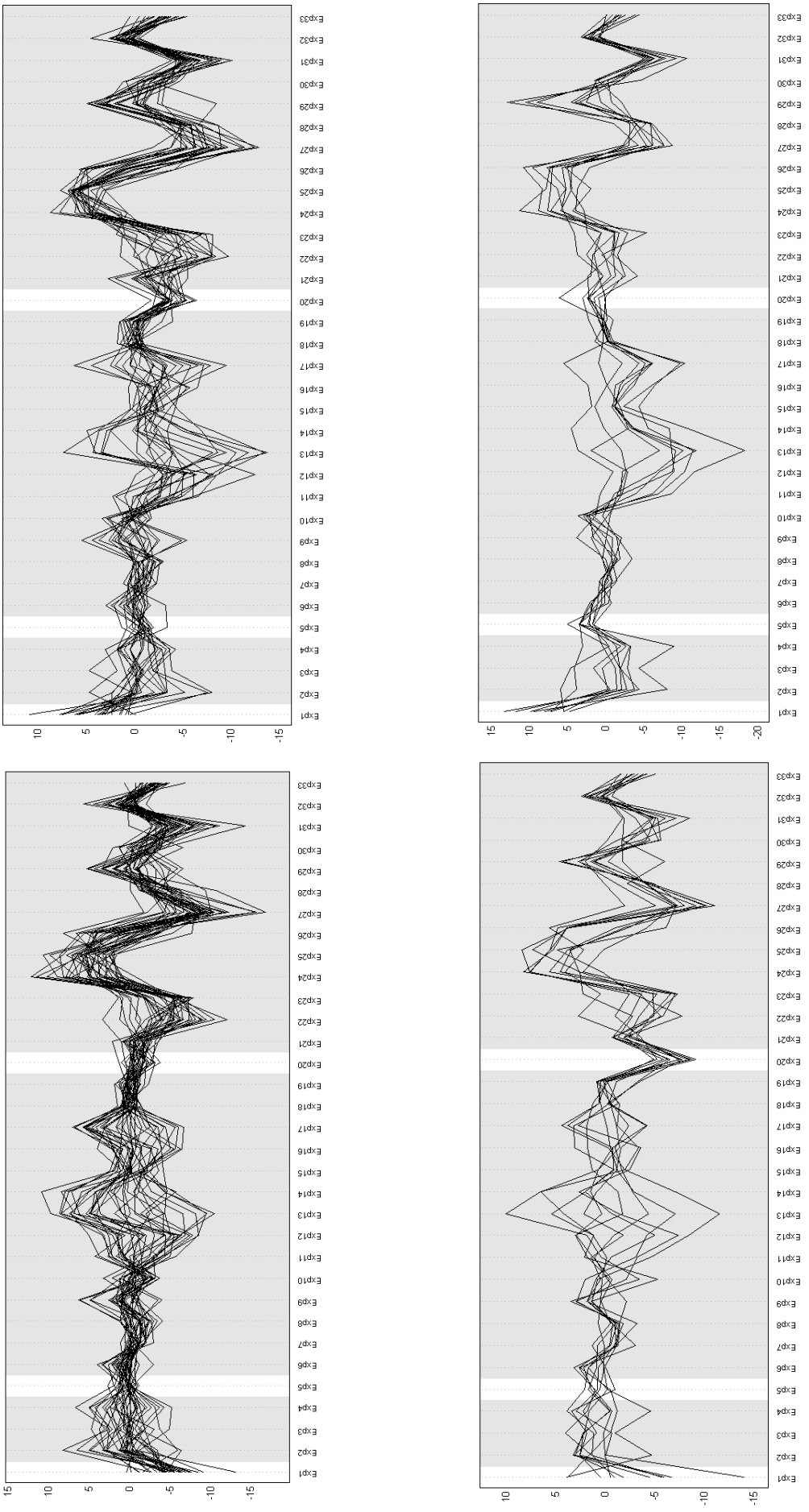

FIG. 3 – Représentation des profils d'expression des gènes de la classe 11 dans la sous-classe 1 (en haut, à gauche), la sous-classe 2 (en haut, à droite), la sous-classe 4 (en bas, à gauche) et la sous-classe 5 (en bas, à droite).

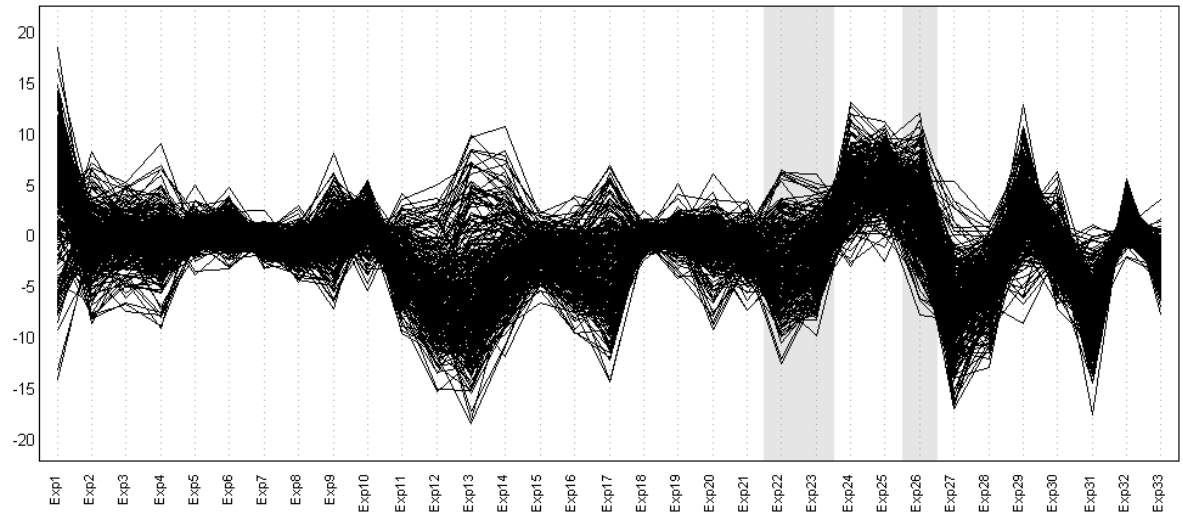

FIG. 4 – Représentation des profils d'expression des 329 gènes de la classe 11.

6 Discussion

La méthode de sélection de variables pour la classification que nous avons illustré sur des données génomiques tire un parti optimal du modèle de mélange gaussien. Elle propose ainsi une typologie du rôle des variables qui permet souvent d'améliorer la pertinence de la classification et vient enrichir systématiquement son interprétation. Elle renforce ainsi l'intérêt de ce modèle de mélange pour la classification qui permettait déjà de sélectionner avec des critères objectifs un nombre de classes et une forme de classification bien fondés (voir par exemple l'article de Biernacki dans ce numéro de la Revue Modulad).

L'application décrite ici montre que cette méthode de sélection de variables en classification constitue un outil intéressant pour émettre des hypothèses précises sur le rôle des gènes et leurs relations.

Le logiciel *SelvarClustIndep* programmé en C^{++} est disponible sur la page web de Cathy Maugis (`http://www.math.u-psud.fr/~maugis`). Il a été adapté à l'URGV pour la prise en compte des données génomiques et permettre les descriptions apparaissant par exemple dans la figure FIG. 1. Il sera à terme disponible dans le logiciel MIXMOD présenté par F. Langrognet dans ce numéro de la Revue Modulad.

Références

AGI : Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408 :796–815.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3) :803–821.

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

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5) :781–793.

Crowe, M. L., Serizet, C., Thareau, V., Aubourg, S., Rouze, P., Hilson, P., Beynon, J., Weisbeek, P., van Hummelen, P., Reymond, P., Paz-Ares, J., Nietfeld, W., and Trick, M. (2003). CATMA : a Complete Arabidopsis GST database. *Nucleic Acids Research*, 31(1) :156–158.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25) :14863–14868.

Gagnot, S., Tamby, J.-P., Martin-Magniette, M.-L., Bitton, F., Taconnat, L., Balzergue, S., Aubourg, S., Renou, J.-P., Lecharny, A., and Brunaud, V. (2008). CATdb : a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic Acids Research*, 36(Database Issues) :986–990.

Gollery, M., Harper, J., Cushman, J., Mittler, T., Girke, T., Zhu, J., Bailey-Serres, J., and Mittler, R. (2006). What makes species unique ? the contribution of proteins with obscure features. *Genome Biology*, 7 :757.

Law, M. H., Figueiredo, M. A. T., and Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9) :1154–1166.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2008). Variable selection in model-based clustering : A general variable role modelling. Technical Report RR-6744, INRIA.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009a). Variable Selection for Clustering with Gaussian Mixture Models. *Biometrics*. To appear.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009b). Variable selection in model-based clustering : A general variable role modeling. *Computational Statistics and Data Analysis*. To appear.

Raftery, A. E. and Dean, N. (2006). Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, 101(473) :168–178.

Samson, F., Brunaud, V., Duchene, S., De Oliveira, Y., Caboche, M., Lecharny, A., and Aubourg, S. (2004). FLAGdb++ : a database for the functional analysis of the Arabidopsis genome. *Nucleic Acids Research*, 32 :347–350.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2) :461–464.